UNIVERSITAT DE BARCELONA

Study of the components that determine the applicability of pathogenicity predictors in the clinical setting

Josu Aguirre Gómez

Study of the components that determine the applicability of pathogenicity predictors in the clinical setting

Doctoral Thesis 2020

Josu Aguirre Gómez

UNIVERSITAT DE BARCELONA

Tesi Doctoral

Programa de Doctorat en Genètica
Facultat de Biologia
Departament de Genètica, Microbiologia i Estadística

Study of the components that determine the applicability of pathogenicity predictors in the clinical setting

Memòria presentada per **Josu Aguirre Gómez** per optar al grau de doctor per la Universitat de Barcelona

Tesi realitzada al Grup de Recerca de Bioinformàtica Clínica i Translacional de Vall d'Hebron Institut de Recerca, sota la direcció del Dr. Xavier de la Cruz Montserrat

Director	**Tutor**	**Doctorando**
Dr. Xavier de la Cruz Montserrat	Dr. Josep Francisco Abril Ferrando	Josu Aguirre Gómez

Barcelona, maig 2020

Ad familiares et amicos.

DECLARATION

I hereby declare I myself carried out the work described in this thesis, except where indicated in the text. The work presented here took place in the group of Clinical and Translational Bioinformatics at the Vall d'Hebron Research Institute under the supervision of the Dr. Xavier de la Cruz Montserrat. Also, I declare that this thesis has not been and will not be submitted in whole or in part to another University for the award of any other degree.

Signed: 

Date: Tuesday, 19 May 2020

Josu Aguirre Gómez

Barcelona

ACKNOWLEDGEMENTS

Mi más profundo y sincero agradecimiento es, en primer lugar, para Xavier; mi director y supervisor en esta tesis, que ha conseguido transmitirme su amor y pasión por la ciencia e infundirme un pensamiento científico y crítico con el que poder afrontar la vida más allá del laboratorio. Me dio la oportunidad de realizar la tesis y me ha ofrecido una orientación inconmensurable a lo largo de estos años; infundiendo en mí un espíritu más curioso, riguroso, crítico, honesto y empático. En definitiva, trabajar y estudiar bajo tu dirección me ha hecho por encima de todo, mejor persona. Has sido el mejor mentor con el que alguien pudiera soñar para realizar su tesis. Por todo ello te estaré eternamente agradecido. Gracias Xavier.

A todos los que han formado y forman parte del grupo de Xavier. A Casandra y Òscar, que han sido unas grandísimas personas y compañeros durante el periodo en el que coincidimos, y de los que tuve la oportunidad de aprender. A Natàlia, la mejor técnica y futura doctora con la que he tenido la suerte de trabajar. Su habilidad para resolver los problemas técnicos y capacidad para explicar los entresijos de la programación no dejan de sorprenderme. A Elena, una trabajadora que no acepta un "es imposible" como respuesta. Tengo la suerte de poder llamaros amigas. A Luz y Selen, a quienes les deseo mucha suerte con sus respectivos doctorados. No cabe duda de que haber formado parte de este grupo ha sido una gran fortuna.

A todos los colaboradores, dentro y fuera del VHIR, que nos han dado la oportunidad de tender puentes y acercar nuestro trabajo. Al grupo de Ramón Martí, a Yolanda y Cora; al equipo de Ricardo Pujol Borrel, en particular a Roger Colobràn. Al equipo de Francesc Palau del hospital Sant Joan de Déu. A todo el equipo que forman Pirepred-Poctefa, a nuestros vecinos más próximos de la UEB.

A los buenos profesores que me han traído hasta aquí, desde la escuela hasta la universidad.

Bereziki zuri Aitor, unibertsitatean zientziarekiko kuriositatea eta exijentzia transmititu zenidalako, eta beti laguntzeko eta entzuteko prest egon zinelako.

Aunque os mencione los últimos, siempre seréis los primeros para mí: familia y amigos. Siempre habéis estado a mi lado apoyándome, mucho antes de que comenzara la tesis. A Ia amoña, que pone empeño a sus 87 años por entender en qué cosiste mi trabajo. Como no podría ser de otra forma, a la "kuadrilla", que siempre están ahí para poner las cosas en perspectiva y hacer que los agobios y dudas pasen a un segundo plano; con vosotros uno se pueda sentir como en casa; en especial a ti Jon. A Laida, Pello, Ander, Miren, Garazi y Andoni, contar con vuestra amistad es un verdadero privilegio. No puedo dejar sin mencionar a mis amigos de Barcelona, mi "kuadrilla catalana", con los que he vivido y compartido todas las alegrías y penurias durante estos años. A Pérez, con quien siempre que conversas aprendes, a Joan, Xec, Manu, y a vosotros, Unai y Luis, con los que además tengo la suerte de compartir piso.

A mi ama, mi aita y Martín.

Sin vuestro apoyo incondicional y entrega (por mucho que haya veces que sea desde la distancia), nunca podría estar donde estoy hoy. Gracias, de corazón.

A todos a los que en un párrafo no os puedo condensar, gracias.

Mila esker.

"The aim of argument, or of discussion, should not be victory but progress."

Karl R. Popper

ABSTRACT

The translation of Next Generation Sequencing (NGS) technologies from the research field to the clinical setting and, specifically, the results obtained in terms of diagnostic yield remain far from expected. This situation is due to our present inability to solve the "variant interpretation problem", which consists in establishing whether a sequence variant is either pathogenic or neural. In this thesis we have focused on how this problem is addressed by pathogenicity predictors, studying the components that determine the applicability of these tools in the clinical setting.

First, we have developed a novel approach to assess pathogenicity predictors in terms of both their performance and their suitability for clinical applications. We present a cost framework for assessing and comparing *in silico* tools, inspired on the use of cost models applied in different fields, from clinical tests to credit assessment in finance. A virtue of this cost framework is that it takes into account the consequences of downstream medical decisions in a simple fashion.

Second, we have studied one of the most important factors limiting the performance of pathogenicity predictors: genetic background. In this part, we have studied the relationship between molecular impact and disease severity in hemophilias A and B, for a specific type of sequence variants: compensated pathogenic deviations (CPDs). We have established, studying a dataset of variants in coagulation factors FVIII and FIX, that the disruptive impact of a mutation is not enough to explain the associated phenotype. In parallel, we have characterized the genetic background of these proteins,

describing at the molecular level its potential to generate phenotypic variability.

Finally, we have characterized the contribution of *in silico* pathogenicity predictors to the variants identified in gene sequencing panels, using as a model a panel designed for Primary Immunodeficiency Disease (PID), developed in the Immunology and Autoinflammatory diseases' groups, at the Vall d'Hebron University Hospital. The results obtained illustrate the limits of *in silico* tools and also a new way to take genetic background into consideration.

RESUMEN

La traslación de las tecnologías de secuenciación de última generación (NGS) del ámbito de la investigación al entorno clínico, y más en concreto, los resultados obtenidos en su rendimiento diagnóstico, continúan lejos de lo esperado. Esta situación se debe a nuestra incapacidad para resolver el "problema de interpretación de las variantes", que consiste en establecer si la variante de una secuencia es patogénica o neutra. En esta tesis nos hemos centrado en cómo se resuelve este problema mediante los predictores de patogenicidad, estudiando los componentes que determinan la aplicabilidad de estas herramientas en el entorno clínico.

En primer lugar, hemos desarrollado una nueva aproximación para evaluar los predictores de patogenicidad en términos de su rendimiento y su idoneidad para aplicaciones clínicas. Presentamos un marco de coste para evaluar y comparar los métodos *in silico*, inspirados en el uso de modelos de coste en diferentes campos, desde los ensayos clínicos hasta la evaluación del crédito en las finanzas. Una virtud de este marco de coste es que contempla las consecuencias de las decisiones médicas finales de una forma sencilla.

En segundo lugar, hemos estudiado uno de los factores más importantes que limitan el rendimiento de los predictores de patogenicidad: el entorno genético. En esta parte, hemos estudiado la relación entre el impacto molecular y la severidad de las hemofilias A y B en unas variantes de secuencia específicas: las variantes patogénicas compensadas (CPD). Estudiando un conjunto de datos de variantes en los factores de coagulación FVIII y FIX, hemos establecido que el impacto disruptivo de una mutación no es suficiente para explicar el fenotipo asociado. En paralelo, hemos

caracterizado el entorno genético de estas proteínas, describiendo a nivel molecular su potencial para generar variabilidad fenotípica.

Finalmente, hemos caracterizado la contribución de los predictores de patogenicidad *in silico* en las variantes identificadas en los paneles génicos de secuenciación, usando como modelo un panel diseñado para la Inmunodeficiencia Primaria (IDP), desarrollado en los grupos de Inmunología y Enfermedades Autoinflamatorias, en el Hospital Universitario Vall d'Hebron. Los resultados obtenidos ilustran las limitaciones de las herramientas *in silico* y también una nueva forma de tener en cuenta el entorno genético.

CONTENTS

1. Introduction: Principles underlying the applicability of pathogenicity predictors in the clinical setting

The research developed in this thesis is devoted to study the components that determine the applicability of *in silico* pathogenicity predictors in the clinical setting. In this introductory chapter, we will describe the current state of the use of Next Generation Sequencing (NGS) experiments in the clinical applications, and how they set the context for the use of pathogenicity predictors. Then, we will focus on the variant interpretation problem and; then, on the main characteristics of pathogenicity predictors, the *in silico* tools designed to address this problem. We will also describe the current performance measures for estimating the success rate (classification/misclassification) of a pathogenicity predictor, and why they provide an incomplete answer to the problem of identifying pathogenicity predictors for clinical applications of NGS. Finally, we will focus on an important component contributing to the applicability limits of *in silico* tools for diagnostic purposes: genetic background.

1.1. NGS and personalized medicine

The release of the first draft sequence of the human genome in 2001 (Craig Venter *et al.*, 2001; Lander *et al.*, 2001) and the completion of the human genome project (Collins *et al.*, 2003) along with the appearance of NGS technologies, increased the hope and expectation towards personalized medicine (Pasche and Absher, 2011; Dammann and Weber, 2012). The rapid development and the increasingly reduced cost of sequencing techniques (Figure 1.1) have led to their generalized application in different areas of medical practice, such as diagnosis, prognosis and therapy (Chen and Snyder, 2013). The purpose of using of these techniques in the clinical setting is to obtain and understand the patient's genome and to take into account the

specific needs and singularities of every patient; in summary, to give them a personalized treatment.

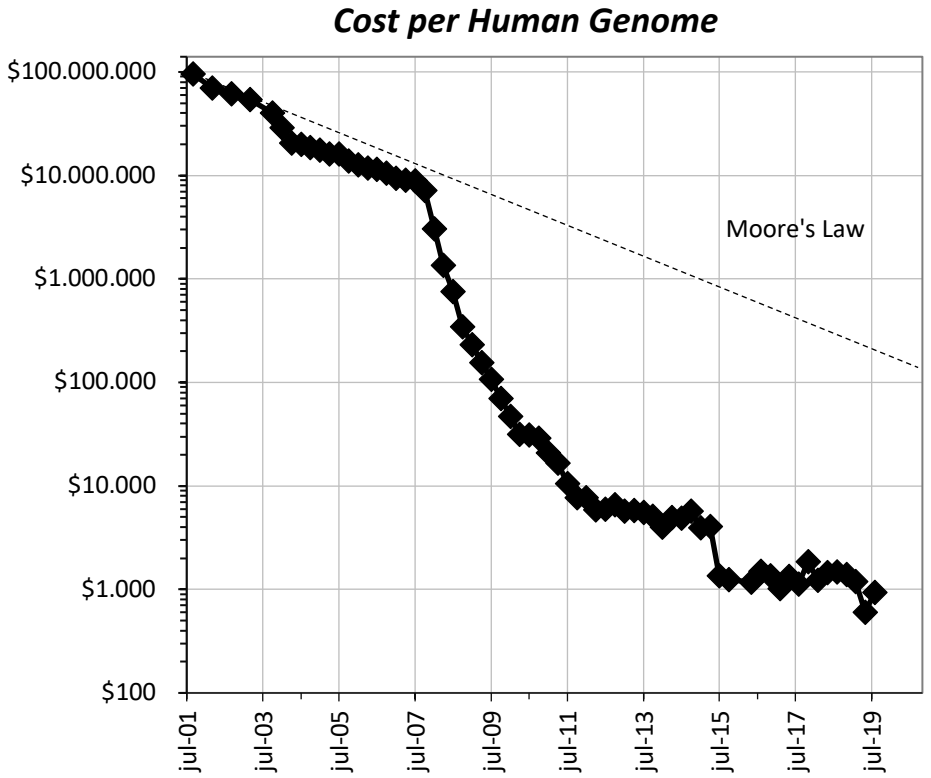


Figure 1.1. Evolution of the sequencing cost of the human genome over time. *Source: www.genome.gov/sequencingcostsdata/ checked on 01-29-2020.*

However, even though there has been a decrease in the cost of NGS technology (Figure 1.1) (Mardis, 2010; Sboner *et al.*, 2011), and although significant steps have been made in understanding the genome and the genetic contributions to human health and disease (Green and Guyer, 2011), the translation of NGS technologies from the research field to the clinical setting (Bertier, Cambon-Thomsen and Joly, 2018) remains far from expected. There are several factors delaying the extended use of NGS in medical applications. For example, the lack of knowledge of the molecular basis of the disease (Hall, Moore and Ritchie, 2016) and the presence of environmental and epigenetic factors that modulate the genotype expression in the

phenotype (Delaney *et al.*, 2016; Han and He, 2016). Also, the design of data processing *in silico* tools is complex, because modelling biological complexity is arduous and mathematically limited (Colijn *et al.*, 2017). In summary, although in the past few years many efforts have been directed to take advantage of the present advances in NGS, e.g., developing guidelines for the evaluations of NGS applications for the diagnosis of genetic disorders (Green *et al.*, 2013; Richards *et al.*, 2015; Matthijs *et al.*, 2016), the promises of a personalized medicine remains unanswered (Roden and Tyndale, 2013), remains unaccomplished.

1.1.1. NGS in clinical diagnosis

NGS enables rapid, cost-effective and, highly accurate genome-scale data generation (Xuan *et al.*, 2013). Its introduction in clinical diagnosis, illustrated in Figure 1.2, represents a considerable advance in our ability to provide diagnoses for patients with rare inherited diseases (Ng *et al.*, 2010; Stranneheim and Wedell, 2016; Kim *et al.*, 2019) or with common but highly heterogeneous disorders (Aspromonte *et al.*, 2019; Marques Matos, Alonso and Leão, 2019; Yska *et al.*, 2019). The idea of using NGS in clinical diagnosis is to go beyond the sequencing of a target-gene or gene groups, to find the maximum number of putative genetic causes of disease, going from single-nucleotide variants (SNVs) to large genomic rearrangements (Chen *et al.*, 2016). Following this strategy, the use of NGS has resulted in improvements in diagnostic yield of ~15-30% (Jamuar and Tan, 2015; Aspromonte *et al.*, 2019; Marques Matos, Alonso and Leão, 2019), respectively. In addition, it has been observed that these diagnostic rates could be increased through reanalysis of the NGS data (Sun *et al.*, 2019). As a consequence, NGS technologies are increasingly used in the clinical practice: they give physicians more resources to better manage their patients, by reducing the number of undiagnosed cases.

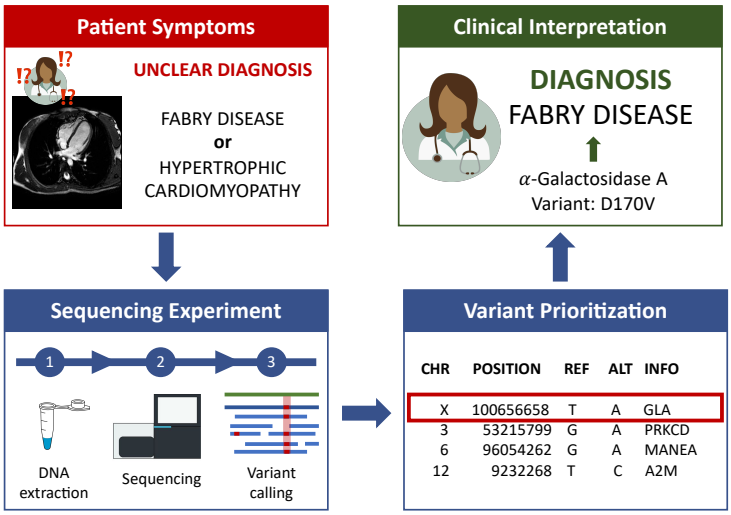


Figure 1.2. Next Generation Sequencing (NGS) application in the clinical setting. *The lack of sufficient clinical evidence to identify a disease is a recurrent problem in medical diagnosis (top-left); the sequencing experiment allows the identification of a sequence variant consistent with one of the possible options considered by physicians (bottom); the use of this information allows establishing the most plausible diagnosis for the patient (top-right).*

The most frequently used applications of NGS in the clinical setting are: whole exome sequencing (WES) and gene sequencing panels (Lohmann and Klein, 2014; Di Resta *et al.*, 2018). It is important to note that these experiments, while really powerful, have some limitations. For example, the confidence and accuracy of variant calling can vary across the genome; in fact, only 990 genes in the genome are found entirely within high-confidence regions. For the subset of clinically relevant genes, like those sorted in (i) ClinVar (Landrum *et al.*, 2014) and OMIM (Amberger *et al.*, 2009), and (ii) the ACMG-reportable (American College of Medical Genetics and Genomics) genes (Olfson *et al.*, 2015), only 74.6% and 82.1% of exonic bases are found in high-confidence regions, respectively (Goldfeder *et al.*, 2016).

Whole exome sequencing (WES) is focused on the coding region of the genome, the exome (Choi *et al.*, 2009; Kim *et al.*, 2010). The

implementation of WES is promising, especially in the case of rare genetic diseases (Gilissen *et al.*, 2011, 2012; Zhang, 2014; Y. Xue *et al.*, 2015; Sawyer *et al.*, 2016), where diagnosis results more arduous for physicians (Nguyen and Charlebois, 2015). Furthermore, WES can identify known disease genes that could not be identified previously due to the genetic heterogeneity associated with the disease (Y. Xue *et al.*, 2015; Sawyer *et al.*, 2016; Di Resta *et al.*, 2018). There are certain clinical scenarios where the 'hypothesis-free' approach makes WES the best diagnostic tool: (i) in extreme genetically heterogeneous disorders where *de novo* variants are the major mutations (O'Roak *et al.*, 2011; Ku *et al.*, 2013); (ii) when two or more unrelated phenotypes are present in the same patient (Cullinane *et al.*, 2011); and (iii) when there is no presence of a key phenotypic feature (Zaidi *et al.*, 2013), being the phenotype irrelevant and the real cause of the disease difficult to identify (Choi *et al.*, 2009; Majewski *et al.*, 2011). In these circumstances, the use of WES will enable to modify a patient's medical management (Belkadi *et al.*, 2015; Baldridge *et al.*, 2017).

As we have previously mentioned, gene sequencing panels are one of the most used NGS applications in clinical diagnosis (Bertier, Cambon-Thomsen and Joly, 2018; Di Resta *et al.*, 2018). These panels are applied to an increasing number of diseases and disorders (BlueShield, 2020), including cardiovascular diseases, neurological disorders, psychiatric conditions, cancer and also reproductive testing, among others. They follow the principle of exome sequencing and cover coding regions of genes (Choi *et al.*, 2009; Kim *et al.*, 2010; Y. Xue *et al.*, 2015; Di Resta *et al.*, 2018), although they also include some exome-adjacent regions (Moret *et al.*, 2019). Gene panels are particularly valuable in: (i) genetically heterogeneous disorders with well-defined disease associated genes (Valencia *et al.*, 2013), (ii) diseases with overlapping phenotypes (for differential diagnosis) (Mook *et al.*, 2013), (iii) diseases sharing certain manifestations but with totally different general

presentations (Lemke *et al.*, 2012), and (iv) disorders sharing genes from a common structure or pathway (Xu *et al.*, 2017).

1.1.2. Issues related to the clinical applicability of NGS technologies

A priori, when a WES assay is used for clinical diagnostic purposes, ~21000 protein coding-genes (Pertea *et al.*, 2018) with all known disease-associated relevant genes (~4600 genes) (Y. Xue *et al.*, 2015) are tested. However, available WES kits do not cover the entire exons of these genes. In fact, they offer a low coverage for many regions of the exome; indeed more than 10% of the whole exome is not covered with the accepted 20x minimum (Y. Xue *et al.*, 2015; Di Resta *et al.*, 2018). In fact, in face of these issues and considering its broader scope, the biomedical community is starting to seriously consider the advantages of using WGS instead of WES. For example, WES needs from two to three times higher sequencing coverage rate than WGS to cover the same amount of bases (Lelieveld *et al.*, 2015). Also, if we compare WGS and WES base-pair coverage and accuracies in coding regions at comparable sequencing depth, we see that WGS is more powerful than WES detecting single nucleotide variants (SNVs) (Biesecker and Green, 2014; Belkadi *et al.*, 2015; Lelieveld *et al.*, 2015). Moreover, although it has limited sensitivity, WGS is more precise than WES detecting structural variants, such as insertions, deletions and translocations (Biesecker and Green, 2014; Belkadi *et al.*, 2015). Besides, the proportion of false-positive variants is higher in WES than in WGS, so WGS systems seem more powerful and efficient to detect potential disease-causing SNVs (Belkadi *et al.*, 2015). However, WGS also presents challenging limitations. For example, the problem of variant interpretation (Figure 1.2, bottom) is more challenging in WGS than in WES, since the number and type of variants detected by WGS is more significant, both quantitatively and qualitatively (Biesecker and Green, 2014; Belkadi *et*

al., 2015) and it also captures intronic and intergenic variants (Biesecker and Green, 2014). In addition, and for the moment, WES is cheaper than WGS, and consequently, more clinical laboratories are using it (Biesecker and Green, 2014).

In recent years, multi-gene sequencing panels have constituted themselves as a powerful alternative to WGS and WES. Globally, gene panels are based on a good knowledge of the biomedical problem if interest, simplifying the interpretation problem present in WGS and WES. Plus, at the technical level, panels ensure a better coverage of the regions of interest (Jones *et al.*, 2013); in addition, they present fewer variants of unknown significance (VUS) – the phenotypic impact of the variant is unknown – improving the diagnostic yield compared to WES (Y. Xue *et al.*, 2015; Di Resta *et al.*, 2018). These crucial advantages over WES and WGS (Group, 2015) are making gene panels preferable in the clinical setting.

The design and composition of gene sequencing panels is not trivial, and it has not been standardized yet, leading to panels that include different numbers of genes for the same or similar clinical conditions, reducing the reproducibility of diagnosis processes. For example, the number of genes included in epilepsy gene panels varies from 70 to 377 (Y. Xue *et al.*, 2015). This variability results from the fact that it is not always clear which genes are associated with a disease, and different authors may have different views of the available evidence. Moreover, to facilitate differential diagnosis, genes that have overlapping phenotypes with the primary disease of the panel must also be included (L. C. Xue *et al.*, 2015; Di Resta *et al.*, 2018; BlueShield, 2020). Finally, it must be mentioned that the composition of panels varies over time (BlueShield, 2020), because of improvements in our knowledge of the genetic cause of disease by the use of NGS technologies (Biesecker and Green, 2014; Belkadi *et al.*, 2015; Lelieveld *et al.*, 2015; Y. Xue *et al.*, 2015; BlueShield, 2020).

Summing up, irrespective of the target region, routine use of NGS in clinical diagnosis requires the following characteristics: high accuracy, simple assays, small and cheap instruments, flexible throughput, short-run times and easy data analysis and interpretation (Desai and Jere, 2012). Moreover, with regard to all the aspects related to information management – like data processing, storage, management and interpretation – NGS also presents enormous challenges (Xuan *et al.*, 2013; De Goede *et al.*, 2016). Many of these requirements still limit the utility of NGS systems in clinical diagnosis (Desai and Jere, 2012; Xuan *et al.*, 2013; Linderman *et al.*, 2014; Lohmann and Klein, 2014; De Goede *et al.*, 2016; Goldfeder *et al.*, 2016). For example, NGS has coverage and accuracies, lower than 100%, resulting in false-positive findings and missing variants (Lohmann and Klein, 2014). However, all the NGS technologies mentioned face a common problem that limits their use in routine clinical diagnosis: the variant interpretation problem. It is posed very simply: once a variant is detected, we have to understand its molecular impact and its relation with the disease. We address this issue in the following section (1.2).

1.2. The variant interpretation problem

As mentioned before, the variant interpretation problem is a key limiting factor for the complete adoption of NGS in routine clinical diagnosis. Why is it hard to establish the effect of sequence variants? That is, why is it difficult to understand how and when a variant will lead to a of its carrier? For many years now, several investigations have focused on the knowledge and understanding of the molecular impact of sequence variation and its relation to disease. A natural approach to address this unsolved question is the use of functional assays, e.g., *in vivo* and *in vitro* experiments (Kitzman *et al.*, 2015; Starita *et al.*, 2017; Bonjoch *et al.*, 2019; Nussinov *et al.*, 2019), metabolic tests, etc. However, currently, there are no functional assays for all genes, and those

performed after the discovery of the variant are both time and resource consuming (Starita *et al.*, 2017). In front of these limitations, the use of computational estimates of a variant's impact, using what are known as *in silico* tools/methods or pathogenicity predictors, appears as a promising alternative to the problem of variant interpretation (Shendure, Findlay and Snyder, 2019).

The computational approach to the variant interpretation problem relies on the type of variants we are aiming at, e.g. single-nucleotide variants (SNVs) in the coding regions or in the non-coding regions of the genome, large insertions/deletions, etc. In this thesis, we will focus on the case of single amino acid variants for two main reasons. On one hand, these variants are known to contribute to disease: they constitute ~58% of the causative variants stored in HGMD (Peter D Stenson *et al.*, 2012). On the other hand, the scientific knowledge required to develop pathogenicity predictors for single amino acid variants has reached an important level of maturity, and comprises different fields of knowledge (Riera, Lois and De la Cruz, 2014), from evolutionary to biophysical studies of proteins, their structure and function, as we will see in the next section.

1.2.1. Characterizing the molecular impact of amino acid variants

Regarding the relationship between a variant and its clinical phenotype, there are two main components: (i) the impact of the variant on the protein sequence, and (ii) the propagation of this impact through the different levels of the biological hierarchy (cell, tissue, organ, system, etc.), a process regulated by the individual's genetic background and environment. The vast majority of pathogenicity methods are based on attributes that only take into account the first component (Sunyaev, 2012), the molecular impact of the variant. These attributes can be divided into two broad groups: (i) those

reflecting stability and structure disruption; and (ii) those measuring disruption of the sequence conservation pattern as represented in multiple sequence alignments (MSAs).

Protein stability (ΔΔG) is a thermodynamic property directly related with the structural basis of the protein function and with its persistence in the cellular environment (Fersht, 1998). Several studies show that this fundamental property is sensitive to sequence mutations (Wang and Moult, 2001; Guerois, Nielsen and Serrano, 2002; Sánchez *et al.*, 2006; Rost, Radivojac and Bromberg, 2016; Ponzoni and Bahar, 2018), and that this is actually the case for pathogenic variants (Ferrer-Costa, Orozco and De La Cruz, 2002; Yue, Li and Moult, 2005; Kucukkal *et al.*, 2015). Some of them because the amino acid change is associated to important changes in the properties of the native amino acid (Ferrer-Costa, Orozco and De La Cruz, 2002), e.g., changes in hydrophobicity, in amino acid volume, etc. In other cases, location in the 3D structure of the affected residue plays a key role (Riera, Lois and de la Cruz, 2014). For example, the impact of mutations on core or surface residues is substantially different. In fact, functionally neutral variants tend to be located at the protein surface (de Beer *et al.*, 2013; Petukh, Kucukkal and Alexov, 2015). However, this is not always the case, because the location of pathogenic variants at the protein surface also suggests that they can disrupt native protein interactions; such as protein-protein, protein-substrate and protein-DNA interactions (Fernández-Recio, 2011; David and Sternberg, 2015; Kucukkal *et al.*, 2015; Blázquez-Bermejo *et al.*, 2019; Navío *et al.*, 2019). In fact, the number of pathogenic variants disrupting these interactions is larger than expected (Sahni *et al.*, 2015) and might be larger than the number of pathogenic variants disrupting protein stability.

It is worth noting that characterizing a variant in terms of 3D-based properties can have the additional advantage of providing a mechanistic understanding of their impact (Figure 1.3). In this sense, it is worth mentioning a new generation of methods based on the use of molecular dynamics simulations, which can provide a deep insight on the structural impact of variants (Angarica, Orozco and Sancho, 2015; Galano-Frutos, García-Cebollada and Sancho, 2019).

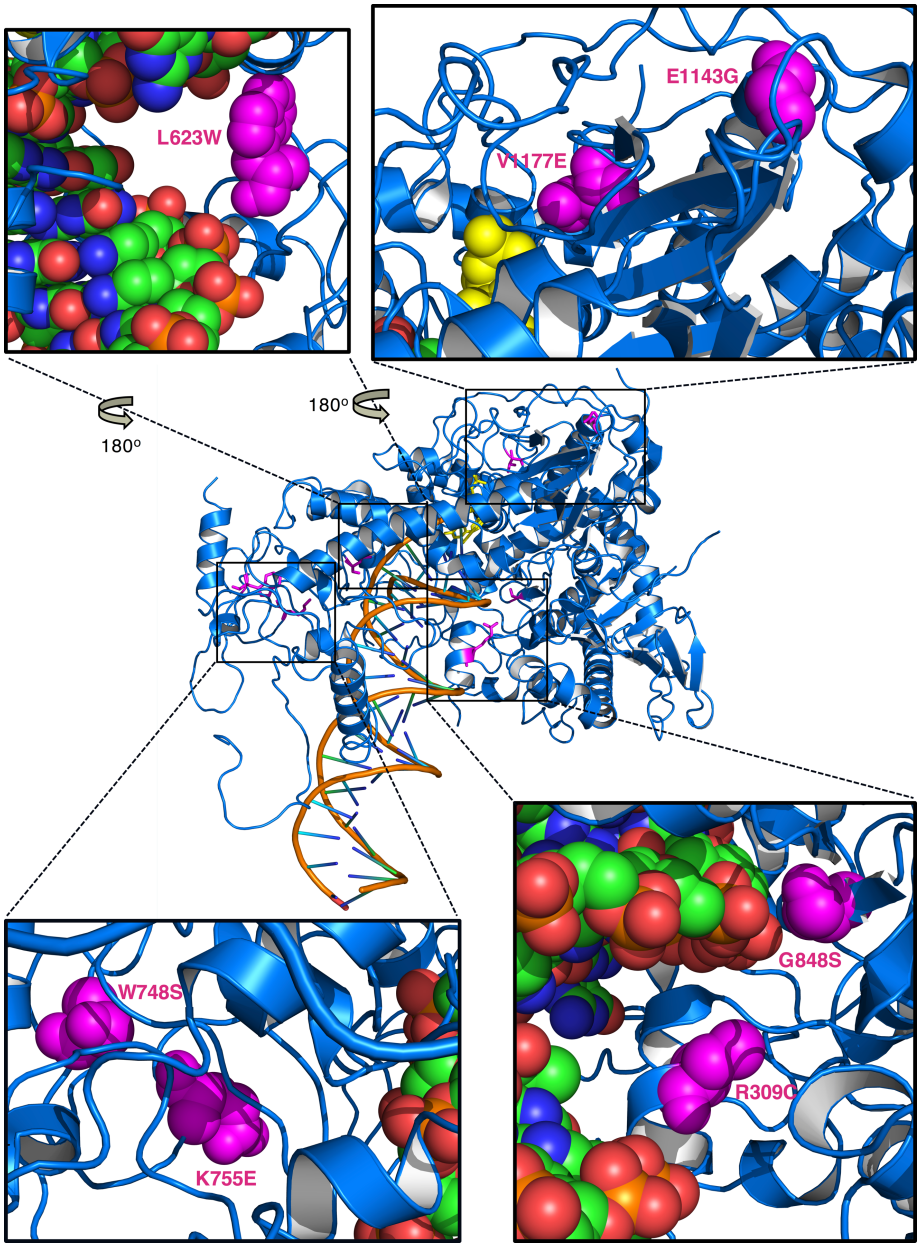


Figure 1.3. Location of the mutated residues within POLG structure predicts different adverse effects on enzyme function. *Structure location of the POLG variants. In the center of the figure, we provide an overall view of the location of all variants in the DNA bound-POLG complex [PDB code: 4ZTZ* (Szymanski *et al.*, 2015)*]. The protein backbone is represented with a blue ribbon; magenta sticks are used for the variants' side-chains; the DNA stretch is shown in orange and the dNTP in yellow. A more detailed view of the specific environment around the variants is provided in the zoomed-in images. Here, sphere representations are used for the variant side-chain (magenta), the DNA*

(red, blue, green and orange) and the dNTP (yellow) molecules. Again, the protein backbone is depicted with a blue ribbon.

The second group of properties utilized to characterize variants is based on the use of MSA: they are measures of the deviation from the conservation pattern introduced by the amino acid replacement. Among these properties we have Shannon's entropy and the position-specific scoring matrix. Shannon's entropy (Cover and Thomas, 2006) is used to estimate the compositional diversity at the locus of the variant in the MSA. It is computed as $-\sum_i p_i \,.\log(p_i)$, where p_i is the frequency of amino acid *i* at the mutation's site. Shannon's entropy varies between 0 and 4.322, with low and high values corresponding to highly and poorly conserved locations, respectively. The second property is $pssm_{nat}$, the value for the native amino acid of the position-specific scoring matrix (Henikoff and Henikoff, 1994), computed as follows: $pssm_{nat} = \log(f_{nat,i}/f_{nat,MSA})$, where $f_{nat,i}$ and $f_{nat,MSA}$ correspond to the frequencies of the native amino acid at the variant locus *i* and in the whole MSA, respectively. It has been shown that these sequence conservation-based features discriminate between neutral and pathogenic variants with a success rate comparable to, or even better than, that of simple structure-based descriptors (Ferrer-Costa, Orozco and de la Cruz, 2004). However, recent studies demonstrate that the predictive value of sequence conservation-based features varies with the protein family (Riera, Padilla and de la Cruz, 2016; López-Ferrando *et al.*, 2017).

A virtue of sequence conservation-based features is that they rely only on the availability of sequence information for the different protein families. As a consequence, in proteins for which we lack structural information (Schwede, 2013) and we cannot apply biophysical methods, we still can use sequence information to discriminate between neutral and pathogenic variants.

Next, we will explain how the fundamental knowledge described in this section is utilized to develop pathogenicity predictors.

1.2.2. The main steps in the development of a pathogenicity predictor

In order to establish the pathogenic nature of a variant, pathogenicity predictors are developed as a two-category classification problem where variants are either pathogenic or neutral and classified using standard machine learning tools (Bishop, 2011; Riera, Lois and De la Cruz, 2014; Hastie, Tibshirani and Friedman, 2017; Camacho *et al.*, 2018). The development of pathogenicity predictors follows the same four standard steps (Figure 1.4): (i) build the variant dataset composed of neutral and pathogenic variants; (ii) select the discriminant properties related to the different functional impact of the mutant amino acid; (iii) choose and train the classification model; and (iv) estimate its performance. Step (ii) is based on our knowledge of the problem we want to solve and has been treated in the previous section (section 1.2.1). In this section we will focus on the three remaining steps.

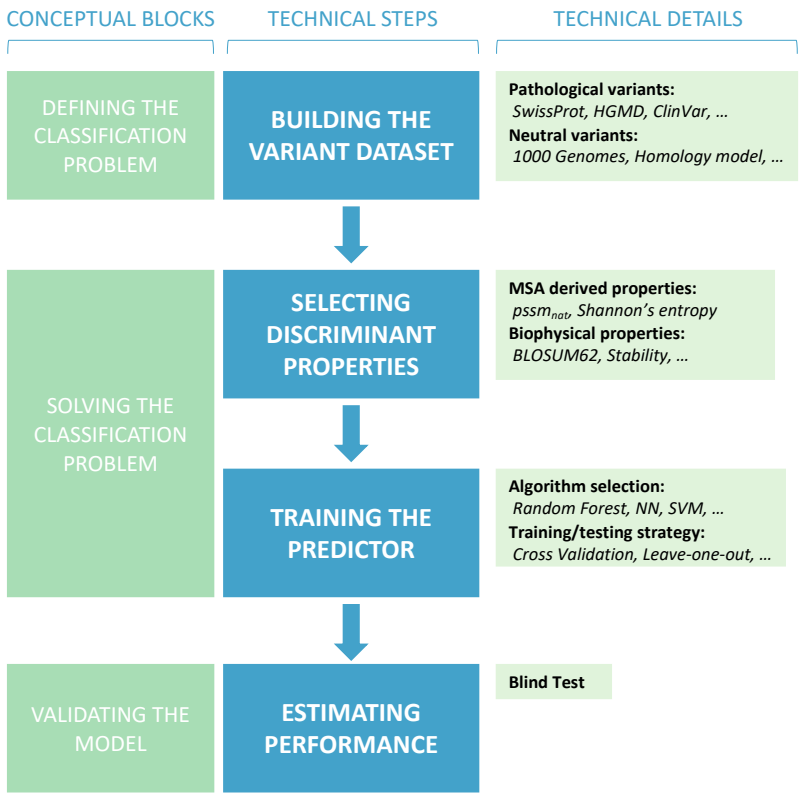


Figure 1.4. The four steps in the development of a method for the prediction of pathogenic variants. *The figure illustrates the correspondence between these steps (blue central boxes) and the main conceptual parts of the prediction problem (green left boxes). The first step is the obtention of the variant dataset, which must include enough pathogenic and neutral variants to reliably estimate all the parameters in the model. Next, we must decide which parameters represent best the relationship between variant impact and disease phenotype. In the third step, the model is produced using one of the many available algorithms (NN, SVM, Random Forests, etc.). Finally, once the model is obtained, its performance must be estimated, so that users can consider to which extent it is adequate for their goals.*

Building the variant dataset. With regard to the training dataset, the pathogenicity predictor learns about the classification task from a collection of known neutral and pathogenic variants. Hence, the prediction goal addressed determines the variant dataset appropriate for the study, and it also defines the applicability range of the resulting pathogenicity predictor. For

example, if the research is focused on a single gene or group of genes, the pathogenicity predictor should be trained by restricting the dataset to variants affecting those genes only (Ferrer-Costa, Orozco and de la Cruz, 2004; Jordan *et al.*, 2011; Crockett *et al.*, 2012; Fechter and Porollo, 2014; Niroula, Urolagin and Vihinen, 2015; Riera, Padilla and de la Cruz, 2016; Shrestha *et al.*, 2018; Padilla *et al.*, 2019). If the predictor is trained with variants associated to a monogenic disorder, it should not be used for scoring variants in complex disorders (Care *et al.*, 2007), and vice versa.

There are different sources from which pathogenic and neutral variants are retrieved to train pathogenicity predictors. Pathogenic variants are obtained from large databases, such as UniProt/SwissProt (Bateman *et al.*, 2017), HGMD (Peter D Stenson *et al.*, 2012), and ClinVar (Landrum *et al.*, 2016) since they are periodically updated and manually curated. However, since there are different annotation and curation protocols, some variants have incorrect pathogenicity annotations (MacArthur *et al.*, 2014), and discrepancies may appear between databases. Neutral variants also can be retrieved from large project databases, e.g., The 1000 Genomes Project (Altshuler *et al.*, 2010) and ExAC/gnomAD (Lek *et al.*, 2016); or from "divergence data". The latter corresponds to those sequence differences between human proteins and their closest homologs (Sunyaev, 2001; Ferrer-Costa, Orozco and de la Cruz, 2004; Bromberg and Rost, 2007; Adzhubei *et al.*, 2010).

The predictive model. Regarding the classification algorithm, classifiers are the technical core of a pathogenicity predictor. These algorithms learn to classify variants into pathogenic or neutral from the instances in the training dataset. There are several machine learning algorithms which are used to build pathogenicity predictors (Bishop, 2011; Hastie, Tibshirani and Friedman, 2017), such as Random Forests and Neural Networks. These

machine learning algorithms are not equally interpretable. Thus, when their performance is similar, we should favor using the most interpretable classifier; for example, using a simple logistic regression instead of using a multilayer neural network. An important aspect that must be considered when choosing the algorithm is the size and composition of the training dataset since small datasets substantially limit the complexity of the predictive models.

Application of the previous steps, in different versions, is at the origin of essentially all known pathogenicity predictors (Pauline C. Ng and Henikoff, 2003; Thomas *et al.*, 2003; Bromberg and Rost, 2007; Bromberg, Yachdav and Rost, 2008; Thusberg, Olatubosun and Vihinen, 2011; Capriotti and Altman, 2011; Sim *et al.*, 2012; De Baets *et al.*, 2012; Adzhubei, Jordan and Sunyaev, 2013; Al-Numair and Martin, 2013; Shihab *et al.*, 2013; Schwarz *et al.*, 2014; Katsonis and Lichtarge, 2014; Niroula, Urolagin and Vihinen, 2015; Tang and Thomas, 2016; Vaser *et al.*, 2016; Ioannidis *et al.*, 2016; López-Ferrando *et al.*, 2017; Rentzsch *et al.*, 2019).

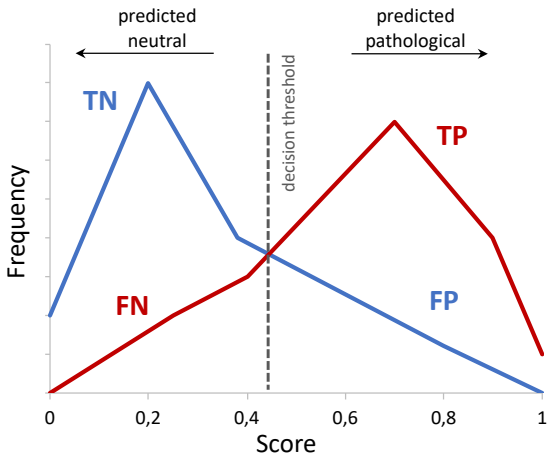
Estimating the predictive performance. Once a predictor has been developed, and before it is delivered to the community of users, an estimate of its predictive performance must be obtained (Witten, Frank and Hall, 2011). This is relevant to establish its suitability for any application, which may have very concrete quality requirements. It is therefore important that the performance parameters employed are meaningful and valuable. We have devoted the next section to this important topic, which is the basis of one of the chapters of this thesis (Chapter 3).

1.3. Measuring the performance of pathogenicity predictors

There are different options to measure the performance of a pathogenicity predictor (Baldi and Brunak, 2001; Vihinen, 2012a). They reflect

the success of the predictor in classifying variants as either neutral or pathogenic. It must be noted that these two discrete classes are obtained from the primary output of *in silico* tools, which is usually a continuous score (normally comprised between 0 and 1) discretized by means of a decision threshold (Figure 1.5).

The performance of a predictor is generally estimated using different combinations of the four elements of the confusion matrix (Figure 1.5) (Baldi *et al.*, 2000): TP (true positives: number of pathogenic variants correctly predicted); TN (true negatives: number of neutral variants correctly predicted as so); FP (false positives: number of neutral variants predicted as pathogenic); FN (false negatives: number of pathogenic variants predicted as neutral). Baldi et al. (Baldi *et al.*, 2000) and Hand et al. (Hand, 2010) catalogue a wide range of performance descriptors for binary classifiers (Baldi *et al.*, 2000; Hand, 2010) that are broadly used in bioinformatics (Ernst *et al.*, 2018; Li *et al.*, 2018; Seifi and Walter, 2018): sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, Matthews correlation coefficient (MCC), and the area under the Receiver Operating Characteristic (ROC) curve (AUC), among others.



		Predicted	
		Pathological	Neutral
Actual	Pathological	TP	FN
	Neutral	FP	TN

Figure 1.5. The output of in silico pathogenicity prediction methods and the confusion matrix. *Most pathogenicity prediction methods present their output as a continuous score, usually between 0 and 1, which is then discretized through a decision threshold into a binary prediction (pathogenic/neutral). Thus, the use of a decision threshold divides the instances into four categories that constitute the four elements of the confusion matrix, shown in the figure as a 2x2 table: TP, FN, TP and TN.*

Sensitivity (also known as True Positive Rate (TPR) or recall) and specificity (also known as False Negative Rate (FNR)) focus on complementary aspects of the prediction problem. Sensitivity measures the ability of a predictor to correctly identify positive cases (pathogenic variants), whereas specificity is the equivalent for negative cases (neutral variants). They are expressed as:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

Positive predictive and negative predictive values, PPV and NPV, respectively, capture the proportion of positive and negative predictions that are truly positive (pathogenic variants) and negative (neutral variants), respectively. They are expressed as:

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

While valuable, these two descriptors depend on the dataset composition, making difficult their use for comparing the work of different authors working with different datasets.

The previous parameters describe separately the performance of the predictor for the two different classes of the problem, neutral and pathogenic. However, when choosing among different predictors it is important to have measures that simultaneously describe the success rate of the predictor for both categories. The most popular among these are accuracy and MCC.

Accuracy is very intuitive, it corresponds to the overall percentage of successful predictions; however, when there is a class imbalance in the mutation dataset (one class is more abundant than the other) it can be misleading (Baldi *et al.*, 2000; Hand, 2010; Kumar *et al.*, 2012; Vihinen, 2013, 2014). Formally, accuracy is expressed as:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

MCC is another performance measure that is highly cited in literature. It is a correlation coefficient, with values comprised between -1 and 1. These two extremes reflect the complete disagreement and agreement in the predictions, respectively; 0 corresponds to a random predictor (Baldi *et al.*,

2000). It is considered more informative for binary classification problems than the previous measures since it considers the four primary quantities in a balanced way (Chicco, 2017). MCC is expressed as:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$$

As we have seen, each of these descriptors have their virtues and defects and, when presenting a new predictor, it is recommended to use several of them to describe its success rate (Baldi *et al.*, 2000; Vihinen, 2012a, 2013).

The previous descriptors are based on the confusion matrix resulting from the use of a specific cut-off value to discretize the problem. However, sometimes we may want to have a more global view of a predictor, independent of the threshold. In this case, AUC, a performance measure computed from the ROC curve, is the most broadly used parameter. The ROC curve is a plot (Figure 1.6) of the proportion of positive cases correctly classified (sensitivity (se), or TPR, on the vertical axis) against that of negative cases incorrectly classified as positives (1- specificity (sp), or FPR, on the horizontal axis). Each point of the curve corresponds to a specific decision threshold (Adams and Hand, 1999). In Figure 1.6, for example, for $Method_A$, for a given decision threshold value, we obtain se=0.76 and sp=0.84. The ROC curve is an increasing function that starts at the bottom left corner of the diagram, which is the origin of the diagram. This point corresponds to a situation in which the method is unable to classify any object as positive and the sensitivity is 0.0. The curve then grows until reaching the top right corner of the diagram, where the method correctly classifies all positive objects as such, and the sensitivity is 1.0. In Figure 1.6 we see the ROC curves of three common methods together with those corresponding to two “special cases”:

(i) a random classifier, which produces a straight diagonal line; and (ii) a perfect classifier, which produces a curve that follows the edges of the ROC square.

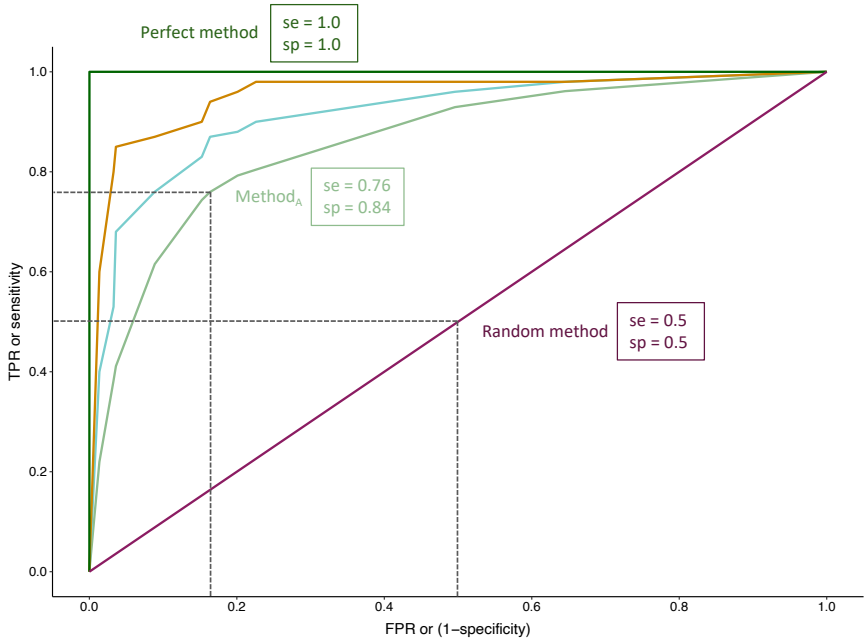


Figure 1.6. The Receiver Operating Characteristic (ROC) curve. *In dark green, following the edges of the graph, we represent the ROC curve of a perfect method, with se=1.0 and sp=1.0. The diagonal, in purple, corresponds to the ROC curve of a random method, with se=0.5 and sp=0.5. We also represent three different ROC curves (one corresponds to Method$_A$, cited in the text, which we identify the point with se=0.76 and sp=0.84). The better the method, the nearer its ROC curve is to that of the perfect method.*

The curves represented in Figure 1.6 are interpreted as follows: the closer is a curve to that of the perfect predictor, the better the corresponding predictor is. However, researchers do not use routinely a visual analysis of closeness; instead they use a value derived from the ROC curve: the Area Under the Curve (AUC) (Figure 1.7). In Figure 1.7 we see that when a classifier outperforms the others, its AUC will be higher. The values of AUC vary between 0 and 1, corresponding to complete disagreement and agreement (perfect method) in the predictions, respectively; 0.5 is the value of a random predictor (Hand, 2009, 2010).

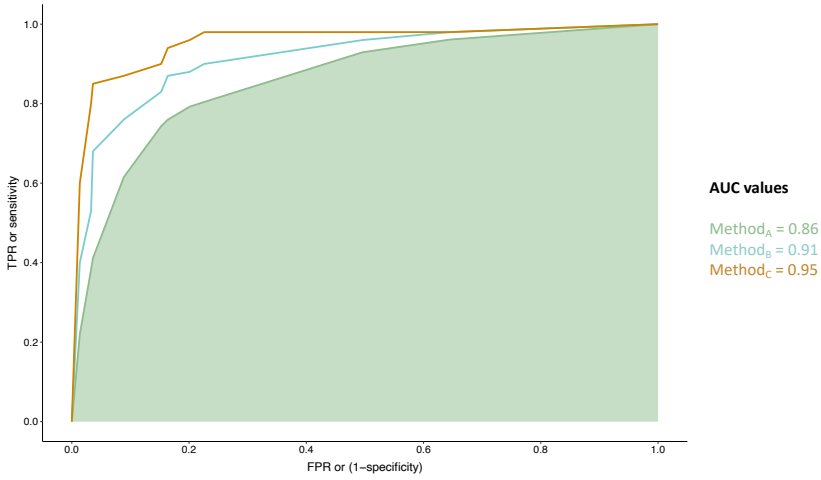


Figure 1.7. The Area Under the Curve (AUC). *The AUC is a performance parameter broadly used to compare classifiers/methods; it corresponds to the area under the ROC curve. Here, the AUC of the Method$_A$, 0.86, is shown in green.*

Although the AUC is objective and it takes into account the four primary/essential quantities (TP, TN, FP and FN), it presents some deficiencies (Hand, 2009, 2010; Hand and Anagnostopoulos, 2013). In particular, a known weakness of the AUC arises when the ROC curves of different methods cross each other, a relatively common situation in pathogenicity predictors (Dong *et al.*, 2015; Grimm *et al.*, 2015; Leong *et al.*, 2015; König, Rainer and Domingues, 2016; Li *et al.*, 2018). This situation is illustrated in Figure 1.8: Method$_A$ and Method$_B$ will be preferable depending on the threshold value. However, one method will have the largest AUC even if the alternative method presents higher/better sensitivity values over most of the range of specificity (Hand, 2010).

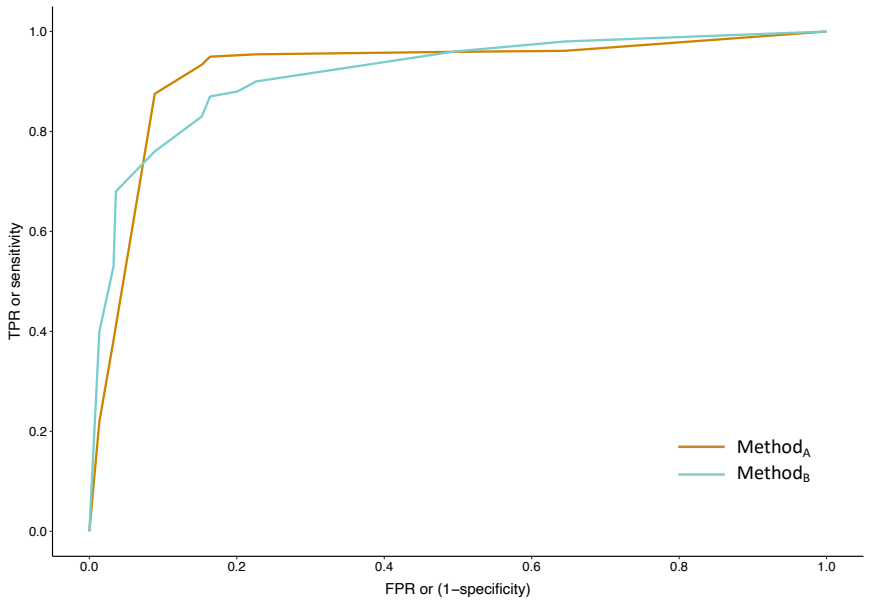


Figure 1.8. Crossing ROC curves. *Here the ROC curves of two different methods (Method$_A$ and Method$_B$) cross each other. Since the ROC curves cross, we cannot use the AUCs to compare the performance of Method$_A$ and Method$_B$* (Hand, 2010).

Other weaknesses of the AUC have been identified by Hand and colleagues (Hand, 2009, 2010, 2012; Hand and Anagnostopoulos, 2013). The most important of these was described in a 2013 paper, where Hand and Anagnostopoulos (Hand and Anagnostopoulos, 2013) showed that "…using the area under the ROC curve is equivalent to evaluating different classifiers using different metrics, and a fundamental tenet of comparative evaluation is that one uses the same measuring instrument on the things being compared: I do not measure your 'size' using a weighing scale calibrated in grams, and mine using a metre rule calibrated in centimetres, and assert that you are 'larger' because your number is greater". That is, using the AUC to compare different methods depends on the nature of the methods compared, and not on the difference between their misclassifications rates (as it should be). Moreover, the AUC may lead to a priori contradictory results with the value, in terms of clinical cost, of the methods compared (Hand, 2009, 2012).

It is of note, however, that with all its pros and cons, the AUC measure is still highly cited (Li *et al.*, 2014, 2018; Mueller *et al.*, 2014; Ghosh, Oak and Plon, 2017; Mahmood *et al.*, 2017), although not always (Rodrigues *et al.*, 2015; Vaser *et al.*, 2016; Kerr *et al.*, 2017; Ernst *et al.*, 2018; Moles-Fernández *et al.*, 2018; Seifi and Walter, 2018).

1.3.1. Limitations of current performance measures

The need for characterizing the success rate of classifiers stems from fact that there are no perfect pathogenicity predictors (López-Ferrando *et al.*, 2017; Mahmood *et al.*, 2017; Li *et al.*, 2018; Seifi and Walter, 2018; Cline *et al.*, 2019), all of them have some degree of misclassification errors (Adams and Hand, 1999; Baldi *et al.*, 2000). Therefore, we cannot use them arbitrarily; we need to identify the one that optimally fills our needs.

The choice of a pathogenicity predictor is in itself a non-trivial problem, since the performance measure used to choose a predictor depends on its target application (Hand and Anagnostopoulos, 2013). Indeed, apart from the above limitations, all the performance measures are independent of the clinical applications, which are characterized by a wide variety of factors, like the cost of large-sequencing experiments, of drugs, of nursing, etc. That is, none of the standard performance measures allows to choose the best pathogenicity predictor, in terms of maximization of the expected clinical utility (Boyko, 1994).

In this thesis we will explore and develop an alternative to standard performance measures (Baldi and Brunak, 2001; Vihinen, 2012a): the cost of a test. The use of cost is frequent in the clinical scenario, where it is used to compare the value of different clinical assays for their routine use by healthcare professionals (Pepe, 2003). Outside medical applications, cost is also broadly utilized, in particular in classification problems from different

fields (Hand, 2010), notably in credit scoring. This has led to the development of some of its formal aspects, like cost curves (Drummond and Holte, 2006) and related measures (Drummond and Holte, 2006; Flach and Matsubara, 2008; Hernández-Orallo, Flach and Ferri, 2013; Hand and Anagnostopoulos, 2019). In general, cost can be a very promising way to characterize pathogenicity predictors in terms of their applicability, because it takes into account simultaneously the performance of the predictor and the specifics of the application context; the latter, in our case, would be the clinical scenario. Cost models are easy to adapt to the case of pathogenicity predictors, because they are devised for two-class problems (Drummond and Holte, 2000, 2006; Hand, 2001). However, this naïve application also has an important limitation: it does not consider that in many cases the output of pathogenic predictors is not binary, it is ternary; there is either a prediction, which is pathogenic or neutral, or no-prediction. This third scenario is important because it means that *in silico* evidence, if used stand-alone, will result in undiagnosed cases that, in the clinic scenario, have an associated cost. Thus, the application of cost concepts for pathogenicity predictor assessment in medical applications should consider both success rate and coverage, because it could then be related to the three possible clinical scenarios: (i) correctly diagnosed, (ii) incorrectly diagnosed and (iii) undiagnosed.

In Chapter 3 of this thesis we address this issue, describing an original cost framework to characterize predictors in terms of their applicability, that integrates the performance of pathogenicity predictors and clinical context.

1.4. The genetic background

As we have seen previously, current pathogenicity predictors do not have a 100% success rate in the identification of pathogenic variants. This may be due to the existence of certain amount of annotation errors in the variants

constituting the training dataset, but most probably on oversimplifications of the predictive model underlying the predictor. In fact, the vast majority of pathogenicity predictors are based on attributes than only take into account the molecular impact of the variant on the protein sequence (Sunyaev, 2012). They ignore that the relationship between a variant and its clinical phenotype is also determined by the propagation of the impact of the variant through the different levels of the biological hierarchy (cell, tissue, organ, system, etc.), a process regulated by the individual's genetic background of the carrier (Badano and Katsanis, 2002; Cooper *et al.*, 2013; Storz, 2016). In particular, the genes belonging to the functional module (or disease pathway) of the mutated gene may play a specifically relevant role in phenotype regulation (Zaghloul and Katsanis, 2010). In this final section of the Introduction, we will focus on the contribution of genetic background in disease, a barely explored field to which we will devote Chapter 4 in this thesis.

Genetic background, which is defined (Gerlai, 2006) as the number of genetic variants carried by and individual in a genomic region, which may go from the whole genome to concrete gene sets, varies across individuals and populations (Hehir-Kwa *et al.*, 2015), and also varies depending on the nature and genes considered. In this context, it is noteworthy the relationship between background and gene size. If we analyze the number of variants in 2504 healthy unrelated individuals (Phase 3 of the 1000 Genomes Project) (Altshuler *et al.*, 2010) in four clinical gene sequencing panels (Illumina's TruSight Sequencing Panels: One, Inherited, Cardio and Autism), we observe a positive correlation between the number of missense variants and gene size (Figure 1.9).

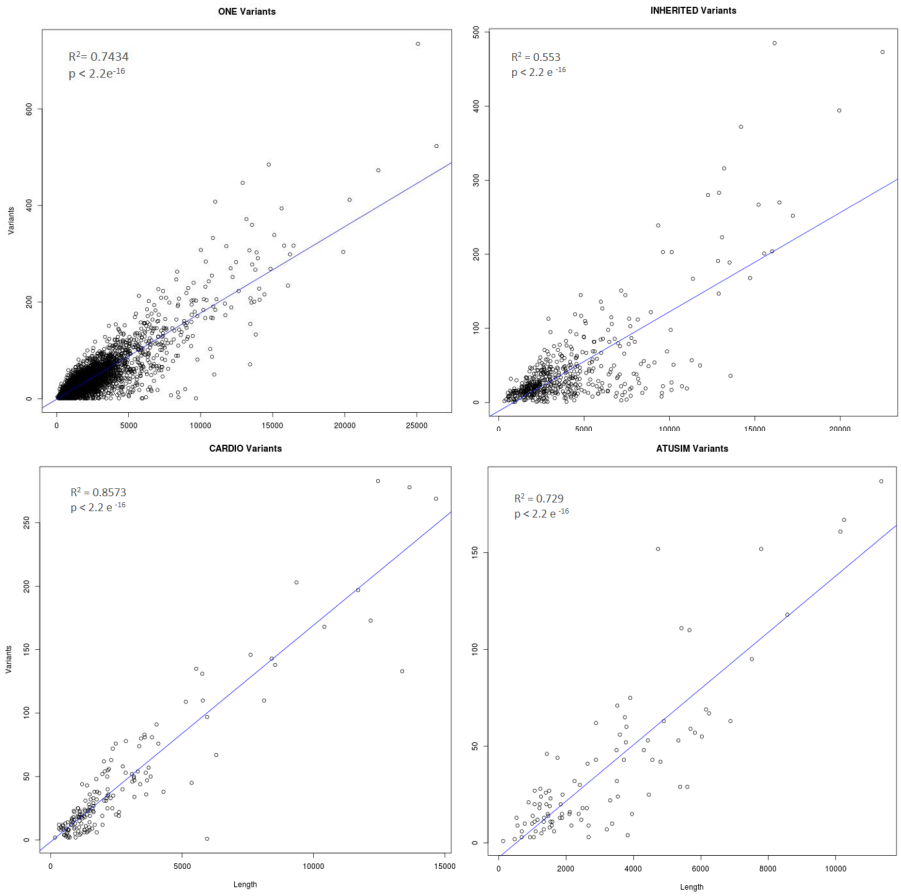


Figure 1.9. Scatterplots representing the positive correlation between gene length and the number of missense variants in four representative clinical gene sequencing panels. *From left to right and top to bottom: (i) correlation in the panel One, (ii) correlation in the Inherited panel, (iii) correlation in the Cardio panel, and (iv) correlation in the Autism panel. The correlation coefficient and p values are present on the top left-hand corner of each panel.*

From a clinical point of view, understanding genetic background is important because it may clarify different aspects of the response of the organism to a pathogenic variant, like the existence of an incomplete penetrance not previously reported (Velasco and Ramírez-Montaño, 2018). However, the clinical impact of the genetic background is not easy to establish; it depends on several factors such as the exact definition of the patient's phenotype (Jamuar *et al.*, 2016), the validation and relevance of molecular

tests, and the appropriate allocation and classification of the putative variants that define the background (Velasco and Ramírez-Montaño, 2018). Regarding this last aspect, several possibilities for further clarification must be considered, e.g., mutational analysis of matched healthy controls, pathogenicity analysis using *in silico* tools, segregation analysis within the family, and the performance of functional assays (Weber *et al.*, 2016).

One of the most intriguing aspects of genetic background is the phenomenon known as "incidentalome", described by Kohane *et al.* (Kohane, Masys and Altman, 2006) as *"… a phenomenon in which multiple abnormal genomic findings are discovered, analogous to the 'incidentalomas' that are often discovered in radiological studies"*. These genomic findings, or variants, are apparently silent, not expressing the disease in the carriers, or at least not apparently. A proper assessment of these findings in the context of screening tests is important because they may act as confusing factors in the diagnosis process. This is particularly true in the case of specific gene panels that contain genes underlying diseases with related phenotypes. From a scientific point of view, the Incidentalome constitutes a paradox that may have multiple explanations, all of them related to one or several possible forms of the suppression of the effect of genetic variants.

1.4.1. Compensated Pathogenic Deviations (CPDs)

The concept of Incidentalome is a recent one, related to large-scale sequencing experiments in the clinical scenario, and it is generally accepted (Kohane, Hsing and Kong, 2012) that it requires some kind of suppressor mechanism. Interestingly, the suppression of the effect of pathogenic variants has been known for several years now, as a result of inter-species and sequence comparisons. As early as 1962, Zuckerkandl and Pauling (Zuckerkandl and Pauling, 1962) described a residue in orangutan hemoglobin that is pathogenic when present in humans, and Motoo Kimura, the father of

the neutral theory of molecular evolution, studied the evolutionary role of these type of variants (Kimura, 1985). Nowadays, we know that about 10% of all the described pathogenic variants in humans are present as native residues in other organisms (Jordan *et al.*, 2015) and, consequently, some compensatory mechanism must be responsible of neutralizing their damaging effect. These variants are known as Compensated Pathogenic Deviations (CPDs) (Alexey S. Kondrashov, Sunyaev and Kondrashov, 2002) and different efforts have been devoted to their study.

Different possible compensatory mechanisms have been described for CPDs, from a specific change in the protein sequence (Barešić *et al.*, 2010) to the accumulation of aminoacidic substitutions that, in an additive manner, make the protein resistant to a theoretical pathogenic variant (Xu and Zhang, 2014; Starr and Thornton, 2016). Moreover, the compensatory effect can be due to the presence of variants in other proteins with which the pathogenic proteins interacts or shares a metabolic pathway (Lehner, 2011). In other words, the compensation of pathogenic variants is an epistatic phenomenon (Lehner, 2011; Starr and Thornton, 2016) in which compensatory variants can appear in *cis* and/or in *trans*.

From a clinical point of view, it is important to identify whether putative pathogenic variants identified in a sequencing experiment are compensated, because if this is the case we can discard them as causative variants. However, establishing whether a variant is compensated or not, that is whether it is a CPD or not, is a difficult task. They have been studied in terms of their molecular impact, severity, and a few defining characteristics are surfacing (Ferrer-Costa, Orozco and Cruz, 2007) that suggest that CPDs would be variants with a milder molecular effect than common pathogenic variants (Barešić *et al.*, 2010).

Identifying CPDs can have another important benefit. The mild molecular nature of these variants suggests that they could be associated with milder versions of the disease. If this were the case, information not only about its potential pathogenic effect, but also about its phenotypic impact (milder or severe), would bring the arrival of precision medicine closer (Niroula and Vihinen, 2016). Indeed, possible pathogenic variants could be stratified into two groups. The first would correspond to variants with severe pathogenic effects, more likely to lead to potentially fatal diseases, thus requiring an immediate action. The second group would correspond to variants whose mild pathogenic effect could result in a milder version of the disease, where it is less urgent to act (Green *et al.*, 2013, 2017).

In Chapter 4 of this thesis we will address this topic, studying the possible effect of genetic background in the case of CPDs found in coagulation factors FVIII and FIX, responsible of hemophilias A and B, respectively.

2. THE OBJECTIVES OF THIS THESIS

OBJECTIVES

The overall goal of this thesis is to advance in the applicability of pathogenicity predictors in the clinical setting, by devising a new form to assess their value and by studying specific systems. In particular, we have focused on the following objectives:

- Obj. 1: Development of an integrative cost framework for analyzing the clinical applicability of pathogenicity predictors. Obtain the formalism and test the results in a system of known predictors.

- Obj. 2: Characterize the relationship between molecular impact and disease phenotype in the context of CPDs, for variants associated to hemophilias A and B. Establish the molecular properties of genetic background for these diseases.

- Obj. 3: Study the application of *in silico* predictors to the characterization of the variants identified in gene sequencing panels, using as a model a panel designed for the diagnosis of patients of Primary Immunodeficiency Disease (PID). Explore a new form to model genetic background.

3. AN INTEGRATIVE FRAMEWORK FOR ANALYZING THE CLINICAL APPLICABILITY OF VARIANT PREDICTIONS

The goal of this chapter is to develop a novel approach to measure the performance of pathogenicity predictors. We present a framework for assessing and comparing the clinical applicability of pathogenicity predictors, inspired on the use of cost models for clinical tests that take into account the consequences of downstream medical decisions.

3.1. Introduction

In spite of early criticisms (Wade, 2010), the contribution of Next Generation Sequencing (NGS) to XXIst century healthcare/medicine is growing steadily (Shendure, Findlay and Snyder, 2019). It has been shown for different genetic conditions that NGS can increase diagnostic yield above 25% (Schwarze *et al.*, 2018). Also, use of NGS is increasingly considered for large-scale screenings (Friedman *et al.*, 2017; Zhang *et al.*, 2019), either alone or in combination with other assays. However, the potential of NGS is still limited by our inability to fully interpret the results of this technique (Starita *et al.*, 2017; Shendure, Findlay and Snyder, 2019): we are still unable to establish with 100% accuracy if the sequence variants identified by this technique are neutral or pathogenic. The consequences of this problem are so severe that it has been referred to as the "variant interpretation catastrophe" (Starita *et al.*, 2017) and the development of algorithms for its solution is considered as one the Grand Challenges of Genomic Medicine (Shendure, Findlay and Snyder, 2019).

This situation has fueled the development of *in silico* pathogenicity predictors, which are computational tools aimed at discriminating between neutral and pathogenic variants. Each of these methods implements a specific algorithm for solving the variant interpretation problem, based on a series of features that reflect the molecular impact of variants (Riera, Lois and de la

Cruz, 2014). Nowadays, several dozens (Riera, Lois and de la Cruz, 2014; Richards *et al.*, 2015; Niroula and Vihinen, 2016; Riera, Padilla and de la Cruz, 2016a; Li *et al.*, 2018; Cline *et al.*, 2019) of these tools are available; however, their performances vary and none of them satisfies all the needs of the healthcare community. For example, the Expert Panels adapting the ACMG-AMP guidelines for variant interpretation to specific diseases (Amendola *et al.*, 2016) may recommend different predictors for each condition. In the case of Hearing Loss, experts recommend (*ClinGen Hearing Loss Expert Panel*, 2018) the use of REVEL (Ioannidis *et al.*, 2016) as a source of computational evidence for applying PP3 and BP4 criteria to missense variants. For the same criteria, in the case of the Li-Fraumeni syndrome, experts recommend (*ClinGen TP53 Expert Panel*, 2019) to combine Align-GVGD (Tavtigian *et al.*, 2006) and BayesDel (Feng, 2017), based on the work of Fortuno et al. (Fortuno *et al.*, 2018); they also explicitly discard the use of SIFT and PolyPhen-2. Interestingly, these two tools are also discarded for CFTR (Dorfman *et al.*, 2010) screenings, but included in screening studies of congenital hypothyroidism (Yu *et al.*, 2018) and metabolic disease (Ko *et al.*, 2018).

In summary, when designing an optimal strategy for applying NGS to a specific healthcare problem, professionals must find the best tool among a large set of options. This is done (*ClinGen Hearing Loss Expert Panel*, 2018; *ClinGen TP53 Expert Panel*, 2019) using success rate measures such as sensitivity and specificity, or Matthews Correlation Coefficient (MCC), which are routinely employed by classifier developers (Baldi *et al.*, 2000; Hernández-Orallo, Flach and Ferri, 2012; Vihinen, 2012b) to gauge the extent to which a tool solves a binary classification problem. Clinical experts use the values of these parameters relating them to the consequences of the medical actions downstream the NGS experiment, to identify which predictor (or combination of predictors) minimizes the risk of medical decision errors. For some intuitive parameters, when considered separately, this process seems easy. For

example, from the point of view of sensitivity, an optimal choice is a method with the highest value of this descriptor, because it will contribute to reduce the chance of leaving patients without treatment. Similarly, for specificity, one would favor the method with the highest specificity, to minimize the risk of treating healthy individuals. The problem is that when selecting pathogenicity predictors, we cannot optimize independently both parameters. In fact, any predictor presents a concrete trade-off between sensitivity and specificity and, except in some obvious cases (Ernst *et al.*, 2018), it is hard to define which one is preferable. The situation becomes more complex if non-intuitive functions of sensitivity and specificity, like the MCC, are employed, e.g., in the adapted guidelines for the Li-Fraumeni syndrome (*ClinGen TP53 Expert Panel,* 2019), because implicit in their values are sensitivity-specificity trade-offs of which we may be unaware. For example, under certain conditions, MCC is symmetric in sensitivity and specificity; that is, one trade-off and its opposite will result in the same MCC. To add further complication, there is no a priori guarantee that the same trade-off, i.e., the same *in silico* tool, will be equally valid across all possible national and international hospital scenarios, given the fluctuations in healthcare practices, involved stakeholders, drug costs, etc.

In this work, we rigorously address this issue and describe a framework that takes into account the two components of the problem: tool performance and clinical context. In our approach, clinical context is encoded using a few cost terms that are associated to three fundamental performance measures –sensitivity, specificity and coverage—to give an integrated, formal model of a predictor. This model allows users to generate a more complete view of the behavior of *in silico* tools in different clinical settings. We illustrate the procedure using a set of sixteen known pathogenicity predictors for missense variants, showing how it provides a clear view of their interrelationships within the clinical context.

3.2. Materials and Methods

In this work, we adapt and extend the general cost framework (Adams and Hand, 1999; Pepe, 2003; Drummond and Holte, 2006; Hernández-Orallo, Flach and Ferri, 2012) to the case of *in silico* tools for pathogenicity prediction. This framework integrates a series of key descriptors of the clinical context and three performance parameters (sensitivity, s_e; specificity, s_p; and coverage, α). The clinical descriptors are described in the Results section, in the context of the framework development. In this section, we explain how we have estimated s_e, s_p and α, following a standard procedure (Vihinen, 2012b; de la Campa, Padilla and de la Cruz, 2017).

We close the Materials and Methods section with the description of some numerical aspects of the computations.

3.2.1. Variant dataset

To estimate the sensitivity, specificity and coverage of each *in silico* tool mentioned in this work, we utilized a set of neutral and pathogenic variants retrieved from the dbNSFP database (Liu *et al.*, 2016), version 4.0b1a, release: December 8, 2018. We imposed three filters. First, the variants should not affect a splicing site. Second, the ClinVar Review Status of the variants should be: “Practice guideline”, or “Expert panel”, or “Criteria provided, multiple submitters, no conflicts”. And, third, we unified the clinical significance classes as follows: “Benign” and “Likely benign” variants were labelled as “neutral”; “Pathogenic” and “Likely pathogenic” variants were labelled as “pathogenic”. At the end of this process, our dataset was constituted by a total of 11093 variants: 3752 pathogenic and 7341 neutral.

3.2.2. Pathogenicity predictors

These tools generate a numerical score that, after comparison with a cutoff value, is utilized to classify a target variant as either neutral or pathogenic. Sometimes, for different reasons (Vihinen, 2020), the predictor does not give a result for a given variant.

In this work, we estimated s_e, s_p and α for a set of sixteen representative pathogenicity predictors: PolyPhen2-HDIV (Adzhubei *et al.*, 2010), PolyPhen2-HVAR (Adzhubei *et al.*, 2010), SIFT (Ng and Henikoff, 2003), CADD (Kircher *et al.*, 2014), MutationTaster (Schwarz *et al.*, 2014a), MutationAssessor (Reva, Antipin and Sander, 2011), REVEL (Ioannidis *et al.*, 2016), FATHMM (Shihab *et al.*, 2013), LRT (Chun and Fay, 2009), PROVEAN (Choi *et al.*, 2012), MetaLR (Dong *et al.*, 2015), MetaSVM (Dong *et al.*, 2015), VEST (Carter *et al.*, 2013), MutPred (Pejaver *et al.*, 2017), Pon-P2 (Niroula, Urolagin and Vihinen, 2015) and PMut (López-Ferrando *et al.*, 2017). For each variant we retrieved the pathogenicity prediction of these tools from the dbNSFP database, except for Pon-P2 and PMut. For these two methods we obtained the predictions from the corresponding website. Then, for each method, we obtained the corresponding s_e, s_p and α applying equations (9.1-9.3) (see Results) directly.

3.2.3. Sensitivity, specificity and coverage

The parameters s_e, s_p and α are characteristic of each *in silico* tool, and are used in the cost models presented here. For a given predictor, they are computed as follows:

$$s_e = \frac{TP}{N_p} \qquad s_p = \frac{TN}{N_n} \qquad \alpha = \frac{N}{N_{tot}}$$

where N_p and N_n are the number of pathogenic and neutral variants, respectively; N (=N_p+N_n) is the total number of annotations generated by the predictor. TP (true positives) and TN (true negatives) are the number of correctly predicted pathogenic and neutral cases, respectively. Finally, N_{tot} is the total number of variants in the dataset. All these parameters were estimated, for each tool, using the predictions for the variants in our variant dataset (see section "Variant dataset"), obtained as described in section "Pathogenicity predictors".

Sensitivity and specificity are related to the two possible misclassification errors of pathogenicity predictors: 1-s_e is the fraction of pathogenic variants classified as neutral, and 1-s_p is the fraction of neutral variants classified as pathogenic.

Coverage is related to a different limitation of pathogenicity predictors: 1-α is the fraction of variants for which the predictor does not provide an answer.

3.2.4. Computations

The cost framework is programmed in Python 3.6. In the following sections we describe the detail of the model's required computations, part of which were done using the package Scikit-learn (Pedregosa *et al.*, 2011).

3.2.5. Numerical computations

The cost models presented in this work involve a certain amount of geometric computations, related to the intersection between straight lines, that may be affected by known inexactness issues (de Berg *et al.*, 2008). To deal with this problem and preserve the geometric coherence of the results, we have added the following ad hoc rules:

- R1. All computations are done using eight digits after the decimal point.

- R2. If two different points are obtained as the intersection between the same two lines, these points will be unified. The coordinates of the resulting point will be the average coordinates of the unified points.

All the computations in the work were done with the package Scikit-learn (Pedregosa *et al.*, 2011).

3.2.6. Computation of the rc_{bd} integral over a polygon region

Our results in Figure 3.10f required computing the value of the integral of the parameter rc_{bd} (Equation (10)) over a polygonal region of the rc_0-rc_1 plane. This double integral was computed following a known approach (Ghorpade and Limaye, 2009). First, we triangulated the polygon, using the package from Scikit-learn (Pedregosa *et al.*, 2011). Second, for each of the resulting triangles, we computed the double integral applying cubature rules (Ghorpade and Limaye, 2009). Finally, we added the results for all the triangles, to obtain the value of the double integral in the polygonal region.

3.3. Results

The results described in this chapter can be grouped in two parts. First, the development of the formalism underlying the cost framework presented and, second, the application of this formalism to a representative set of predictors. The formalism presented takes concepts from Decision Theory, Geometry and Computer Science, and it is constituted by a series of lemmas and propositions that underly the algorithmic version of the framework and the resulting scripting. The application to sixteen predictors, puts the cost

framework to practice, showing how integrating methods' performances and clinical setting provides a radically different view from the standard approach to the selection of *in silico* predictors, based on performance only.

Before starting with the formal part, in this paragraph we advance an intuitive view of our approach can be obtained by considering the application of *in silico* tools in a medical context. In this context, the evidence provided by pathogenicity predictors may: contribute to establish (i) a correct or (ii) an incorrect decision on the pathogenic nature of the variant (misclassification error), or (iii) be inconclusive. Along the text, we will refer to these three possibilities as situations (i), (ii) and (iii). In the medical context, a cost is associated to each of them. In situation (i), the correct decision may contribute to the accurate diagnosis of a patient's condition, thus channeling him/her to the adequate treatment procedure. No out-of-budget costs are incurred. However, situations (ii) and (iii) will either cause damage or delay the accurate diagnostic, thus augmenting the burden of patients and their families. From an economic point of view, situations (ii) and (iii) generate additional/unexpected costs that affect both institutional and family budgets. Reducing these costs (to which we will refer as budget deviations) is a main goal when choosing annotation tools for clinical applications (Pepe, 2003).

3.3.1. The cost framework

Here we describe the cost formalism derived for this work. Its starting point is the expression of the average misclassification cost in classification problems (Adams and Hand, 1999; Drummond and Holte, 2006). In the binary case, where a classifier is trained to discriminate between two classes, 0 and 1, the average misclassification cost, c_{mi}, is expressed as a function of (Adams and Hand, 1999): f_0 and f_1, the probabilities of misclassifying classes 0 and 1,

respectively; the corresponding misclassification costs, c_0 and c_1, and the probabilities, π_0 and π_1, that an object comes from class 0 or 1, respectively:

(1) $c_{mi} = \pi_0 f_o c_0 + \pi_1 f_1 c_1$

This formalism is general and can be applied to any problem where assessing the cost of using a classifier is relevant. In our case, we will apply it to the use of *in silico* tools for identifying pathogenic variants in sequencing experiments carried in a clinical setting. These tools, also referred as pathogenicity predictors, generate a numerical score that, after comparison with a cutoff value, is utilized to assign the class of a target variant. In this context, we write c_{mi} as:

(2) $c_{mi} = \rho(1 - s_e)c_0 + (1 - \rho)(1 - s_p)c_1$

where classes 0 and 1 correspond to pathogenic and neutral variants, respectively. c_0 and c_1 are misclassification costs. c_0 is the cost ensuing from annotating pathogenic variants as neutral, and c_1 is the cost ensuing from annotating neutral variants as pathogenic. ρ is the proportion of pathogenic variants in the genome region sequenced (single gene, gene panel, whole exome, or whole genome). Comprised between 0 and 1, the value of ρ is generally unknown, although it will vary with the size and location of the sequenced region and with ethnicity (Auton *et al.*, 2015; Marín, Aguirre and de la Cruz, 2019). Finally, s_e and s_p are the sensitivity and specificity of the annotation method/pathogenicity predictor, respectively. These method-specific parameters measure the success rate of a tool in predicting pathogenic and neutral variants, respectively (Vihinen, 2012b).

3.3.1.a. Model of budget deviations when prediction coverage is 100%

Expression (2) reflects the costs in unwanted deviations from the budget assigned to medical processes when classifiers produce an output, and

we will refer to them as "budget deviations". In this scenario the average misclassification cost (c_{mi}) is equal to the average budget deviations cost (c_{bd}).

(3) $c_{bd} = \rho(1 - s_e)c_0 + (1 - \rho)\big(1 - s_p\big)c_1$

A priori, c_{bd}, can be used to compare pathogenicity predictors in terms of cost. However, this is difficult because the values of c_0 and c_1, which correspond to different cost scenarios, depend on a wide variety of factors related to the different healthcare stakeholders, and are hard to estimate (Adams and Hand, 1999). Therefore, instead of using c_{bd} directly, we can utilize its normalized version $rc_{bd} = c_{bd}/c_T$, where $c_T = c_0+c_1$ is the "cost magnitude" (Hernández-Orallo, Flach and Ferri, 2012), thus preserving the power of cost models to identify the most adequate tool for a given context. To this end, we rewrite (3) as:

(4) $c_{bd} = c_T[\rho(1 - s_e)rc_0 + (1 - \rho)\big(1 - s_p\big)rc_1]$

where $c_T = c_0+c_1$, and $rc_i = c_i/c_T$ (i=0,1). The values of the rc_i are normalized and their sum is equal to 1:

(5) $rc_0 + rc_1 = 1$

We define rc_{bd}, the relative average budget deviation, as:

(6) $rc_{bd} = \frac{c_{bd}}{c_T} = \rho(1 - s_e)rc_0 + (1 - \rho)\big(1 - s_p\big)rc_1$

From (6), we see that there is a monotonic relationship between c_{bd} and rc_{bd}. Consequently, we can use rc_{bd} to compare pathogenicity predictors with the same results as when using c_{bd}. That is, when a method is preferable over another in terms of c_{bd}, then it is also preferable in terms of rc_{bd}, and vice versa.

To simplify the use of rc_{bd} we replace rc_0 by 1-rc_1 (from (5)). It allows us to reduce in one the number of parameters in (6), and we obtain, after some reordering:

(7) $$rc_{bd} = [(1-\rho)(1-s_p) - \rho(1-s_e)]rc_1 + \rho(1-s_e)$$

The values of rc_{bd} are comprised between 0 and 1. It is important to note, that (7) is the equation of a line in which rc_1 is the independent variable. The range of rc_1 values is comprised between 0 and 1, the open unit interval to which we will refer as $\mathcal{I}$. $\mathcal{I}$ comprises all possible clinical cost scenarios resulting from a concrete combination of misclassifications errors when only misclassification costs are considered (situation (ii) above). This is because rc_0 and rc_1 are normalized versions of the misclassification costs c_0 and c_1, respectively. Therefore, rc_0=1-rc_1 (from (5)) and we only need one parameter to define the cost scenarios.

rc_{bd} is a parameter that quantitatively illustrates how inaccuracies in a method's performance translate to medical consequences (measured in terms of cost) under a specific clinical context. It can be used to compare pathogenicity predictors with the same result as using c_{bd}, because these quantities are monotonically related.

We can use (7) to compare different classifiers, across different clinical scenarios. In this comparison, our goal is to divide the range of rc_1 values in intervals such that a specific method prevails over the others in each of them. To show how this can be done, let us start with the case of two methods, M_1 and M_2, with rc_{bd} values $rc_{bd}(1)$ and $rc_{bd}(2)$, respectively. The values of rc_1 solving the equation $rc_{bd}(1) = rc_{bd}(2)$ define the boundaries between the regions for which M_1 is preferable to M_2 ($rc_{bd}(1) < rc_{bd}(2)$) and M_2 is preferable to M_1 ($rc_{bd}(2) < rc_{bd}(1)$). Here, the solution of this equation is a single value,

$rc_{1,eq}$, that divides $\mathcal{I}$ in the desired intervals (0, $rc_{1,eq}$) and ($rc_{1,eq}$, 1). Which method is preferable in each interval can be easily established comparing $rc_{bd}(1)$ and $rc_{bd}(2)$ in the midpoint of the intervals. It may happen that one of the methods prevails over the other across the whole (0,1) interval, e.g. when $rc_{1,eq} < 0$.

Comparison of more than two methods to identify which one is more beneficial to our purposes, and under which circumstances, is a common situation when devising general strategies for annotating protein variants in a clinical context. For example, in a 2015 article on the guidelines for variant interpretation (Zhang *et al.*, 2019) the ACMG-AMP presents healthcare professionals with a list which, although incomplete, has 16 methods for annotating missense variants. To generalize the previous model to M_i (i=1,N) methods we, first, need to find the intersections points, p_{ij}, between all possible pairs of rc_{bd} lines; second, we sort them. Then, the resulting set of intervals (0, p_{ij}), (p_{ij}, p_{kr}),…,(p_{st}, 1), is explored, and pairs of adjacent intervals for which the same method is preferable, are unified. The final list of intervals constitutes the desired distribution of methods across the range of rc_1.

3.3.1.b. Model of budget deviations when prediction coverage is not 100%

Expression (2) is adequate when classifiers produce an output; however, this is not always the case. In fact, the proportion of cases for which a predictor produces an output, known as coverage of the predictor, may vary substantially between methods (de la Campa, Padilla and de la Cruz, 2017; Vihinen, 2020). Unclassified variants are not a misclassification error, but also have economical and non-economical consequences in a clinical context because they reduce the information available for medical decisions. Particularly, when these depend mostly on the output of the sequencing

experiment. Indeed, unclassified variants may delay a diagnosis or increase its uncertainty, with the consequent burden on the patient and families, increase in the number of requested tests, etc. To take into account this effect, we have extended the model in (2), to include also the contribution of incomplete coverage. All these costs together result in unwanted deviations from the budget assigned to medical processes, and we will refer to them as "budget deviations". The average budget deviation, c_{bd}, resulting from a given annotation tool can be obtained treating misclassification errors and lack-of-coverage as independent, as described in the probabilistic tree diagram in Figure 3.1.

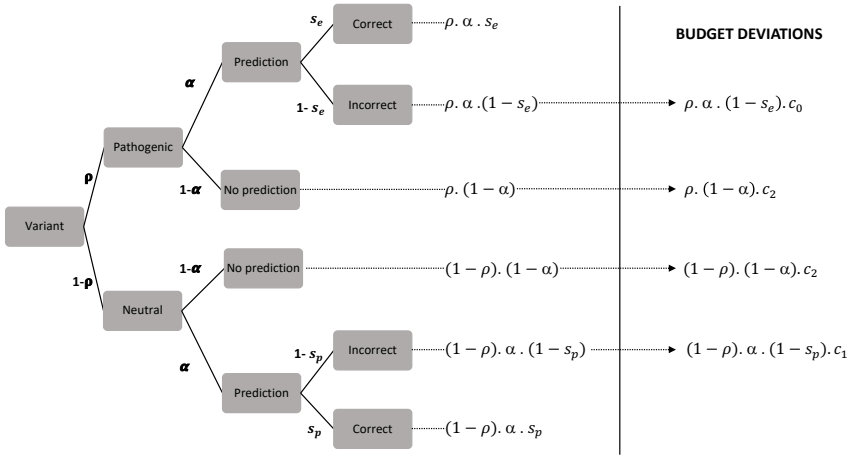


Figure 3.1. Probabilistic tree diagram underlying the cost framework presented in this work. *Each branch corresponds to a different situation and the corresponding probability is written by its side. Multiplying probabilities along branches gives the probability of an event that is the combination of different situations. For example, a pathogenic variant can be incorrectly predicted as neutral; the probability of this event is: $\rho\alpha(1-s_e)$. Separated by a vertical line, we find the corresponding budget deviations, which, after addition and reordering, give equation (8).*

Multiplying the probabilities along the branches and adding the results at the terminating nodes, after some simple reordering, gives:

(8) $c_{bd} = \alpha\rho(1 - s_e)c_0 + \alpha(1 - \rho)\big(1 - s_p\big)c_1 + (1 - \alpha)c_2$

where α is the coverage of the method; c_2 is the cost resulting from not classifying a variant; ρ, c_0, c_1, s_e and s_p have been described when introducing (2).

The parameters s_e, s_p and α are specific of the annotation method. They can be estimated from a sample of known pathogenic and neutral variants as follows:

(9.1) $s_e = \frac{TP}{N_p}$ (9.2) $s_p = \frac{TN}{N_n}$ (9.3) $\alpha = \frac{N}{N_{tot}}$

where N_p and N_n are the number of pathogenic and neutral variants, respectively; $N(=N_p+N_n)$ is the total number of annotations generated by the pathogenicity predictor. TP (true positives) and TN (true negatives) are the number of correctly predicted pathogenic and neutral cases, respectively. Finally, N_{tot} is the total number of variants in the sample.

The values of c_0, c_1 and c_2, which correspond to different cost scenarios, are much harder to estimate (Adams and Hand, 1999), because they depend on a variety of factors related to the different stakeholders. However, we can use normalized values instead (Adams and Hand, 1999; Drummond and Holte, 2006). To this end, we rewrite (8) as:

(10) $c_{bd} = c_T[\alpha\rho(1 - s_e)rc_0 + \alpha(1 - \rho)\big(1 - s_p\big)rc_1 + (1 - \alpha)rc_2]$

where $c_T=c_0+c_1+c_2$, and $rc_i=c_i/c_T$ (i=0,2). The values of the rc_i are normalized and their sum is equal to 1:

(11) $rc_0 + rc_1 + rc_2 = 1$

We define rc_{bd}, the relative average budget deviation, as:

(12) $rc_{bd} = \frac{c_{bd}}{c_T} = \alpha\rho(1 - s_e)rc_0 + \alpha(1 - \rho)\big(1 - s_p\big)rc_1 + (1 - \alpha)rc_2$

From (12), we see that there is a monotonic relationship between c_{bd} and rc_{bd}. Consequently, we can use rc_{bd} to compare classifiers instead of c_{bd} with the same results.

We can reduce the number of parameters in (12) by replacing rc_2 by 1-rc_0-rc_1 (from (11)). We obtain, after some reordering:

(13) $rc_{bd} = [\alpha\rho(1 - s_e) + \alpha - 1]rc_0 + \big[\alpha(1 - \rho)\big(1 - s_p\big) + \alpha - 1\big]rc_1 + 1 - \alpha$

The domain of rc_{bd} in terms of rc_0 and rc_1 is a triangular region in the rc_0-rc_1 plane such that rc_0 and rc_1 are comprised between 0 and 1, and rc_0+rc_1<1. We will refer to this triangle as $\mathcal{T}$.

It is important to note that the points in $\mathcal{T}$ (the pairs (rc_0, rc_1) $\in \mathcal{T}$) reflect the relative costs of the different failure types (misclassification errors, failure to annotate a variant) associated to the use of an annotation tool in the clinical context. This is relevant when using (13) to compare methods in terms of cost. In this comparison, our goal is to identify the cost scenarios, or (rc_0, rc_1) pairs, for which one method is preferable over others. In practical terms, we will aim at identifying the region in $\mathcal{T}$ where the rc_{bd} of one method is always lower than that of the others.

To address this problem, let us consider two methods, M_1 and M_2, and their respective rc_{bd} values, $rc_{bd}(1)$ and $rc_{bd}(2)$, according to (13). The points for which $rc_{bd}(1)$= $rc_{bd}(2)$, form a line, l_{12}, in the rc_0-rc_1 plane that has for equation:

(14)$\{\rho\big[\alpha_1\big(1 - s_{e,1}\big) - \alpha_2\big(1 - s_{e,2}\big)\big] + \alpha_1 - \alpha_2\}rc_0 + \{(1 - \rho)\big[\alpha_1\big(1 - s_{p,1}\big) - \alpha_2\big(1 - s_{p,2}\big)\big] + \alpha_1 - \alpha_2\}rc_1 + \alpha_2 - \alpha_1 = 0$

where $s_{e,i}$, $s_{p,i}$ and α_i (i=1,2) are the method-specific sensitivity, specificity and coverage.

The line ℓ_{12} may cross $\mathcal{T}$, dividing it into two convex polygons (Figure 3.2). One of them is constituted by the cost scenarios for which $rc_{bd}(1) < rc_{bd}(2)$; that is, for which M_1 is preferable to M_2. The second polygon will correspond to the opposite situation, $rc_{bd}(1) > rc_{bd}(2)$, for which M_2 is preferable to M_1. If ℓ_{12} does not cross $\mathcal{T}$, then only one method will always have the lowest rc_{bd} value within the triangle and, consequently, will be preferable in all cost scenarios.

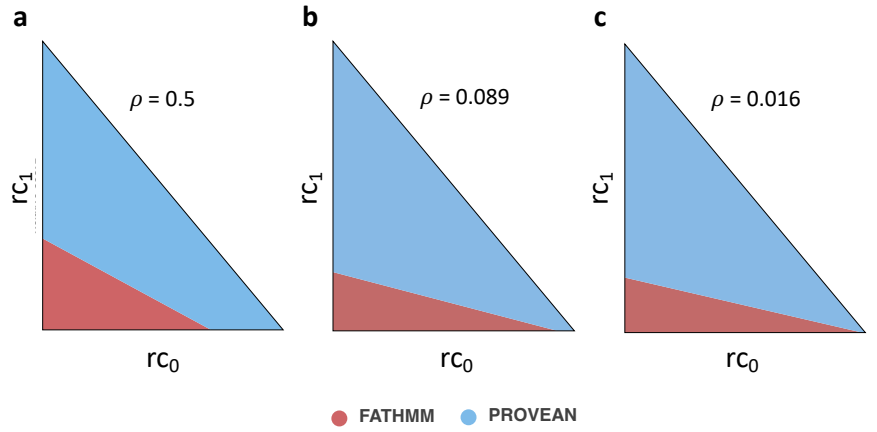


Figure 3.2. Comparison of two methods (FATHMM and PROVEAN) using the cost framework. *In a) we see how the line defined by equation (14) divides $\mathcal{T}$ into two polygons, one where FATHMM prevails (dark red) and another where PROVEAN prevails (blue). The figure also shows, a)-c), how different values of ρ (the fraction of pathogenic variants in the sample) can modulate this result.*

From (14), we can see that the division of $\mathcal{T}$ by ℓ_{12} depends on ρ: different values of this parameter may change the amount of cost scenarios assigned to each method (Figure 3.2). In the following we will concentrate in showing how the comparison between methods can be extended from two to

N; that is, we will address the problem of dividing $\mathcal{T}'$ when more than two methods are available. In this development, we will keep the value of ρ fixed. Subsequently, we will explore how different values of affect this result, when applying the cost framework to a set of sixteen pathogenicity predictors.

When considering N (N>2) methods, the goal will be to identify, for each method, the cost scenarios, or $\mathcal{T}'$ region, for which this method is preferable over the others. To attain this goal, we need a procedure that divides $\mathcal{T}'$ into m regions, r_k (k=1, m), with only a unique method per region with the lowest rc_{bd}.

The next three sub-sections (3.3.2-3.3.4) are devoted to prove, first, that we can establish the r_k regions by generalizing to N methods the approach described for two methods, and, second, to describe how we can actually obtain the r_k regions using a computational procedure based on the standard Breadth-First Search (BFS) algorithm. The starting point will be $\mathcal{L}_N=\{l_{ij}\}$ (i=1, N-1; j=i+1, N), the set of lines resulting from all the possible pair comparisons between methods using rc_{bd}. In the next section (3.3.2), we prove that $\mathcal{L}_N$ lines traversing $\mathcal{T}'$ divide this triangle into a set $\mathcal{P}_N=\{p_l\}$ of convex polygons (Proposition 1), and that within each polygon only one of the N methods will have the lowest rc_{bd} value (Proposition 2). In the short section 3.3.3, we describe how we can identify the method with the lowest rc_{bd} per $\mathcal{P}_N$ polygon. Finally, in section 3.3.4, we will describe how by modeling our problem as a graph problem (Propositions 3 and 4) we can apply an adapted version of the BFS algorithm to find the elements of $\mathcal{P}_N$. The r_k regions are then obtained as the union of all the polygons in $\mathcal{P}_N$ associated to a concrete method.

3.3.2. Division of the cost triangle into a set of convex polygons by the $\mathcal{L}_N$ lines

Note. In our proofs in this section and other sections, we use several results about convex polygons that can be easily found in the books of Lee (Lee, 2012) and of Yaglom and Botyanskii (Yaglom and Boltyanskii, 1961). The most important ones are explicitly cited, providing page and theorem number.

As we have seen before, the triangle $\mathcal{T}$ is the domain of the rc_{bd} function associated to each method, and the points within the triangle correspond to different cost scenarios. Here, we show that $\mathcal{T}$ can be divided into m regions (Figure 3.10a), r_k (k=1, m), such that there is only a unique method per region with the lowest rc_{bd}. These regions are constituted by a set of convex polygons that divide $\mathcal{T}$ and are obtained after performing all the pairwise comparisons between available methods. In this section, we establish some key results required to identify these regions.

Our starting point is a set of N methods that we want to compare in terms of rc_{bd}. From the previous section, we know that there is a line associated to each comparison. After performing all the possible M=N(N-1)/2 pairwise comparisons between methods, we obtain a set $\mathcal{L}_N=\{l_{ij}\}$ (i=1, N-1; j=i+1, N) of lines. These lines cut $\mathcal{T}$ producing a division of this triangle into a set $\mathcal{P}_N$ of convex polygons (Proposition 1).

Proposition 1. Let $N\epsilon\mathbb{N}$ be an arbitrary number of methods compared using rc_{bd} (Equation (13)) and let $\mathcal{L}_N=\{l_{ij}\}$ (i=1, N-1; j=i+1, N) be the set of lines resulting from all the pairwise comparisons between the methods. These lines cut $\mathcal{T}$ (the triangular domain of rc_{bd}) into a set of $\mathcal{P}_N=\{p_l\}$ of convex polygons.

Proof. By induction

Base case. For N=2, there will be only one line in $\mathcal{L}_N$ since there is only one comparison between two methods. When the line contains no interior point of $\mathcal{T}'$, either because it does not intersect with $\mathcal{T}'$, or because it is a supporting line of it, $\mathcal{P}_2$ will have a single element, $\mathcal{T}'$, which is convex because it is a triangle. If the line contains at least a point interior to $\mathcal{T}'$, then it will cut $\mathcal{T}'$ in exactly two points (Yaglom and Boltyanskii, 1961). The line segment uniting these two points is a chord of the polygon (Lee, 2012) and, by the "Polygon Splitting Theorem" (Lee, 2012), divides $\mathcal{T}'$ into two convex polygons.

Induction step. Here we show that if the proposition is true for N, then it is true for N+1. That is, we want to show that if $\mathcal{L}_N$ divides $\mathcal{T}'$ into a set of convex polygons $\mathcal{P}_N$, then $\mathcal{L}_{N+1}$ divides $\mathcal{T}'$ into a new set of convex polygons that we will call $\mathcal{P}_{N+1}$.

We know that the set of lines resulting from the comparison of N+1 methods, $\mathcal{L}_{N+1}$, will contain the lines corresponding to the comparisons between the first N methods, $\mathcal{L}_N$, and between these N methods and an additional N+1-th method, $\{l_{i,N+1}\}_{i=1,\,N}$, that is:

(15) $\mathcal{L}_{N+1} = \mathcal{L}_N \cup \{l_{j,N+1}\}_{i=1,N}$

Cutting $\mathcal{T}'$ with the lines in $\mathcal{L}_{N+1}$ is equivalent to cutting it with the lines in $\mathcal{L}_N$ and then with those in $\{l_{i,N+1}\}_{i=1,\,N}$, since order is irrelevant to the final result. Therefore, $\mathcal{P}_{N+1}$ will be the result of cutting the polygons in $\mathcal{P}_N$ with the lines in $\{l_{i,N+1}\}_{i=1,\,N}$. When we cut $\mathcal{P}_N$ with the first line, $l_{1,N+1}$, we will create a

new division of $\mathcal{T}$ in which each of the polygons split by $l_{1,N+1}$ will be replaced by two children polygons (i.e. no polygon traversed by a line belongs to the new division of $\mathcal{T}$). Next, we will repeat this process for the remaining lines in $\{l_{i,N+1}\}_{i=1,\,N}$ until we obtain $\mathcal{P}_{N+1}$. At the end of each step, the division of $\mathcal{T}$ will be constituted by the set of $\mathcal{P}_N$ polygons unaffected by the $l_{1,N+1}$ line (these polygons are convex because the proposition holds for N), and by the children of the affected polygons. Given that the affected polygons are convex (again because the proposition holds for N), the children will also be convex, by the "Polygon Splitting Theorem" (Lee, 2012). Therefore, at the end of each step, the resulting division of $\mathcal{T}$ will be constituted by a set of convex polygons and, consequently, $\mathcal{P}_{N+1}$, which is obtained at the end of the final step, will be constituted by convex polygons only. □

By construction, the edges of the polygons in $\mathcal{P}_N$ are noncollinear line segments that belong either to the $\mathcal{L}_N$ lines or to the original segments defining $\mathcal{T}$. The vertices of the polygons in $\mathcal{P}_N$ can be: the $\mathcal{T}$ vertices (in few cases), intersection points between the $\mathcal{L}_N$ lines, and intersection points between these lines and the triangle edges.

A simple consequence of the construction process described is Lemma 1, which will be subsequently used in the demonstration of Proposition 3.

Lemma 1. No polygon $p \in \mathcal{P}_N$ has an interior point from any polygon $q \in \mathcal{P}_N$.

Proof. Let us assume that there exists such a polygon p, then there will exist a polygon $q \in \mathcal{P}_N$ such that at least one edge from q cuts p. The $\mathcal{L}_N$ line

corresponding to this edge will then cut p, which is in contradiction with the procedure utilized to generate $\mathcal{P}_N$ (in which any polygon traversed by a line from $\mathcal{L}_N$ is replaced by the two children polygons). □

When comparing N methods in terms of rc_{bd} we will use the partition $\mathcal{P}_N$ to build the regions r_k within which only a single method has the lowest rc_{bd}. First, however, we will show that within each polygon in $\mathcal{P}_N$ only one method has the lowest rc_{bd} (Proposition 2).

Proposition 2. Let $\mathcal{P}_N$ be the set of convex polygons obtained after dividing $\mathcal{T}'$ using $\mathcal{L}_N$, the set of lines associated to the pair comparisons between N methods. There is only one method with the lowest cost within each polygon in $\mathcal{P}_N$.

Proof. Let us assume that the proposition is not true, and that there is a polygon $p \in \mathcal{P}_N$ within which two methods, m_i and m_j, have the minimal rc_{bd} value at points q_i and q_j, respectively.

By construction of $\mathcal{P}_N$ (see Proposition 1), p is not traversed by any $\mathcal{L}_N$ line. Therefore, the $l_{i,j}$ line cannot pass between q_i and q_j, and both points must lie on the same half-plane relative to $l_{i,j}$. This is in contradiction with the fact that in this half-plane only one method, either m_i or m_j, can have the lowest rc_{bd} value (see above), not the two of them. □

The previous proof is easily generalizable to the case of k methods with minimal rc_{bd} values with the same polygon, by successively considering pairs of methods.

Once $\mathcal{P}_N$ is obtained, we can proceed to group the polygons for which the same method has the lowest rc_{bd} value and build the r_k regions (Figure 3.3) as:

(16) $r_k = \bigcup_{i \in \Omega_k} p_i$

where Ω_k are the indexes of the p_i polygons in $\mathcal{P}_N$ for which the k-th method has the lowest rc_{bd} value. We will refer to $\{r_k\}$ (k=1, m), the division of $\mathcal{T}$, as $\mathcal{R}_N$.

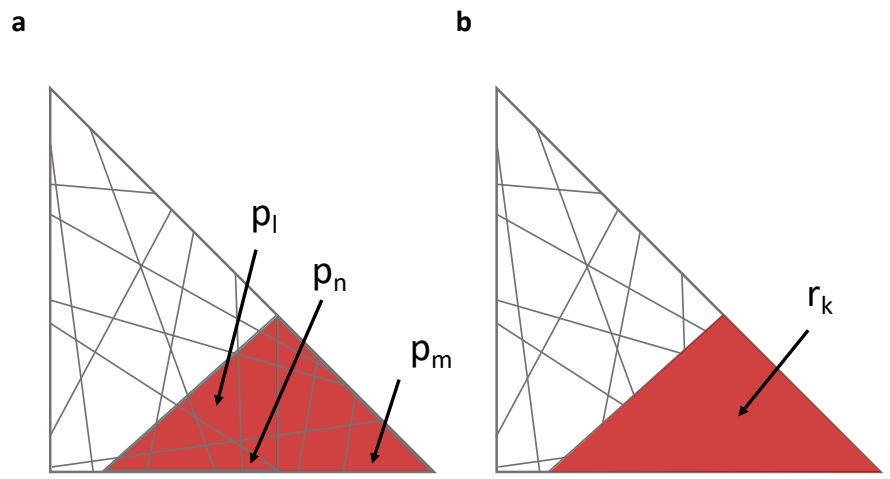


Figure 3.3. Polygon unification. *Unification of the polygons in which a specific method has the lowest average cost (**a),** shown in red) results in a more simplified view **b)** of the regions assigned to each predictor by our approach.*

3.3.3. Obtention of the method with the lowest rc_{bd} within each polygon

We follow a three-step procedure. First, for each polygon in $\mathcal{P}_N$ we compute the average point of its vertices, which lies within the polygon because it is convex. Second, we obtain the rc_{bd} value at this average point for the N methods. And third, we sort the methods according to rc_{bd} and choose

the method with the lowest value. By Proposition 2, this method is unique for the polygon considered.

3.3.4. Building a set of convex polygons $\mathcal{P}_N$ using Breadth-First Search (BFS)

Here we describe the procedure followed to obtain all the polygons in $\mathcal{P}_N$, a step required to obtain the r_k regions (Equation (16)). We give the precise steps and develop the formal aspects required for modeling this task as a graph problem that can be practically solved using the BFS algorithm. The procedure is valid for any arbitrary N, although in this work it is used for N=16. It is constituted by the following four steps.

FIRST STEP. Obtain the set $\mathcal{L}_N=\{l_{ij}\}$ (i=1, N-1; j=i+1, N) of lines corresponding to all the pair comparisons between the N methods using rc_{bd} (Equation (14)). As explained before, these lines cut $\mathcal{T}$ producing a division of this triangle into a set $\mathcal{P}_N$ of convex polygons (Proposition 1). The vertices of these polygons are the intersection points between the $\mathcal{L}_N$ lines, and the edges correspond to the line segments between noncollinear, consecutive vertices.

SECOND STEP. Build the sets of vertices (VP) and edges (EP) of the polygons in $\mathcal{P}_N$. For VP, we first compute the intersections between the lines in $\mathcal{L}_N$, keeping only the points falling inside $\mathcal{T}$ or at its boundaries. These points are included in VP. Then, we compute the intersection between the lines in $\mathcal{L}_N$ and the boundaries of $\mathcal{T}$. The resulting points are added to VP. Finally, we include in VP the three vertices of $\mathcal{T}$. The set EP is constituted by all the $\overline{v_i v_j}$ line segments joining two consecutive vertices v_i, v_j (v_i, $v_j \in$ VP) in a $\mathcal{L}_N$ line; note that no vertex is allowed between v_i, v_j (Figure 3.4). Note also that every

segment in EP corresponds to the edge of a $\mathcal{P}_N$ polygon, as shown in the following lemma.

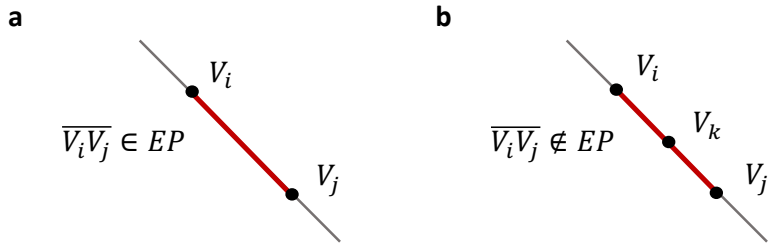


Figure 3.4. a) Allowed, and b) non-allowed situations for the elements of EP.

Lemma 2. Any segment in EP is the edge of at least one polygon in $\mathcal{P}_N$.

Proof. We will assume that there exists a segment $\overline{v_iv_j} \in$ EP which is not the edge of any $\mathcal{P}_N$ polygon.

If $\overline{v_iv_j}$ has points interior to $\mathcal{T}$, these points will be interior to at least one $\mathcal{P}_N$ polygon. Therefore, the line passing through this segment will cut this polygon (or polygons in case there is more than one). This is impossible by construction of $\mathcal{P}_N$ (in building $\mathcal{P}_N$ any polygon traversed by a line from $\mathcal{L}_N$ is replaced by the two children polygons).

Let us now consider that $\overline{v_iv_j}$ belongs to one or several of the edges of $\mathcal{T}$. If $\overline{v_iv_j}$ is included within an edge $\overline{v_rv_s}$ of a single polygon in $\mathcal{P}_N$, then, we would have that $v_r*v_i*v_j*v_s$, which is impossible for the elements of EP which are made of adjacent vertices only (Figure 3.4). If $\overline{v_iv_j}$ points are distributed between the edges of different polygons from $\mathcal{P}_N$, then at least one vertex of these polygons would be comprised between v_i and v_j. This is impossible for the elements of EP (Figure 3.4). In summary, both situations lead to a

contradiction and $\overline{v_i v_j}$ cannot belong to either one or several of the edges of $\mathcal{T}$. □

THIRD STEP. For each $v_i \in VP$ we identify all the $\overline{v_i v_j}$ segments meeting at this point (Figure 3.5a). Then, for each $\overline{v_i v_j}$, $\overline{v_i v_k}$ pair forming a consecutive angle (i.e. an angle formed by three consecutive vertices, Figure 5.3b) we will find the only convex polygon with edges corresponding to segments in EP and having no interior points from other EP segments. This polygon will be added to $\mathcal{P}_N$.

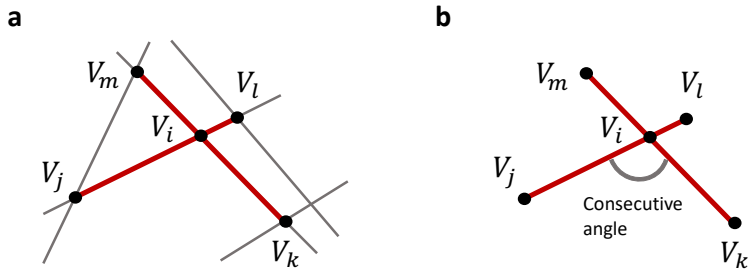


Figure 3.5. a) Segments meeting at a given point. b) Illustration of the consecutive angle.

This strategy is based on the fact that all polygons in $\mathcal{P}_N$ are unique in the sense of the following proposition.

Proposition 3. Let $p \in \mathcal{P}_N$ a polygon with two edges $\overline{v_i v_j}$, $\overline{v_i v_k}$. There exists no convex polygon q, different from p, with two of its edges being $\overline{v_i v_j}$ and $\overline{v_i v_k}$, its remaining edges belonging to EP, and no interior point from a polygon in $\mathcal{P}_N$.

Proof. We will consider that q exists and explore the different possibilities that may arise.

Because p ∩ q = $\{\overline{v_i v_j}, \overline{v_i v_k}\}$ and both polygons are convex, then Int p ∩ Int q ≠ ∅ and:

If p = q, the proposition is proved.

If p ≠ q, there are three options. If all the points in q are interior to p, then the set of edges from q that join vertices v_i and v_k (Figure 3.6a) will be interior to p. Because, these edges belong to EP, they then necessarily belong to polygons in $\mathcal{P}_N$ (by Lemma 2). This is in contradiction with the fact that p does not have interior points from other $\mathcal{P}_N$ polygons (by Lemma 1). We reach an equivalent contradiction in the case when all the points in p are interior to q (Figure 3.6b). And, again, we reach the same contradiction when both p and q have interior points from the other (in this case, the points may come from full segments or fragments of segments) (Figure 3.6c). □

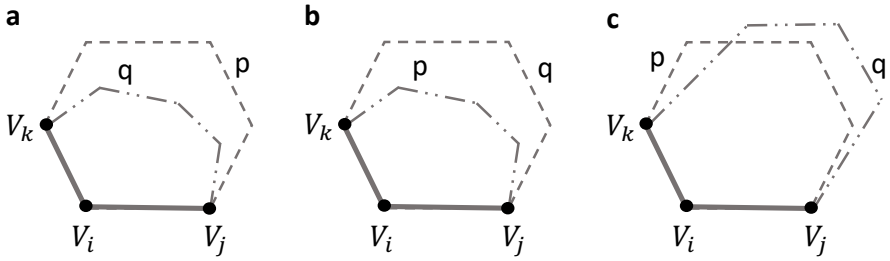


Figure 3.6. Different situations for the relative location of polygons p and q. a) *All the points in q are interior to p;* ***b)*** *all the points in p are interior to q; and* ***c),*** *both p and q have interior points from the other.*

FOURTH STEP. To obtain the list of polygons that constitute $\mathcal{P}_N$ we will enumerate all the convex polygons having v_i as one of their vertices, for i=1, |VP|. To this end, we model this task as a graph problem and use BFS to find the polygons contributed by each vertex. Our starting point is the unweighted,

undirected graph G(V, E), in which V and E correspond to VP and EP, respectively. In this graph, any polygon in $\mathcal{P}_N$ will have a corresponding cycle.

In our procedure, for each $v_i \in V$, we use BFS to obtain the shortest cycle associated to every pair of adjacent edges, $\overline{v_i v_j}$ and $\overline{v_i v_k}$, forming a consecutive angle (Figure 3.5b). This shortest cycle corresponds to a polygon $\mathcal{P}_N$ (as shown in Proposition 4, below). Some conditions are added to the BFS, to adapt it to our problem:

- C1. A cycle cannot have more than one segment from the same line, to guarantee the convexity of the polygons associated to the shortest cycles.

- C2. A cycle cannot have repeated vertices, to guarantee the convexity of the polygons associated to the shortest cycles.

- C3. To guarantee that for each edge we enumerate all the polygons sharing it, every edge has a counter that is decreased by one each time it is included in a cycle. Once the counter reaches zero, the edge will be excluded from future searches. The starting value of the counter will vary depending on the edge's origin. It will be equal to 1 when the edge belongs to one of the sides of the triangle; it will be equal to 2 when the edge belongs to one of the lines in $\mathcal{L}_N$.

- C4. To guarantee that for each vertex we enumerate all the polygons sharing it, every vertex has a counter that is decreased by one each time it is included in a cycle. Once the counter reaches zero, the vertex will be excluded from future searches. The starting value of the counter will vary

depending on the number and origin of the edges that include the vertex (Figure 3.7).

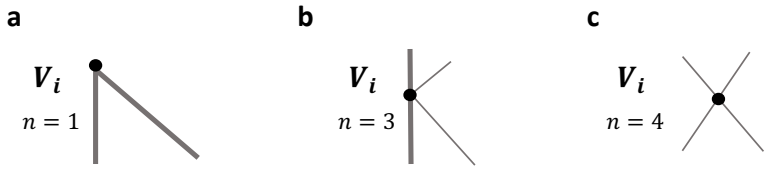


Figure 3.7. Initial values of the counter of each vertex. *In the three figures, thick grey lines correspond to the edges of the triangle $\mathcal{T}'$ (the cost domain, see section 5.3.1), and thin grey lines correspond to the lines dividing $\mathcal{T}'$ and associated with the different pair comparisons between methods.* ***a)*** *The vertex is one of the three triangle vertices.* ***b)*** *The vertex is the intersection between a triangle edge and the line associated with the comparison between two methods.* ***c)*** *The vertex is the intersection between the lines associated to two different pair comparisons between methods.*

- C5. Once a minimal cycle is found, it is excluded from the search.

- C6. For every minimal cycle found, we check that the corresponding polygon does not have any of the remaining vertices in VP as an interior point. This situation may arise, due to numerical inaccuracies, in regions of $\mathcal{T}'$ very dense in vertices; the affected polygon was excluded from the final list of polygons. No significant impact in the subsequent calculations was found.

In the following, and to end this section, we show Lemma 3 and Proposition 4, which are key results in which we establish that each shortest cycle found with the BFS corresponds to a polygon in $\mathcal{P}_N$. That is, the underly the applicability of our cost framework, as expressed in equations (12)-(14).

Lemma 3. Let c_p be a minimal cycle in G(V, E) found with our adapted version of BFS, and passing through the vertex v_i. Then, p, the corresponding polygon, has no interior points from another polygon in $\mathcal{P}_N$.

Proof. We will show that assuming that p has an interior point A, belonging to VP, leads to contradiction.

If p has an interior point, A, belonging to VP, then A will be the intersection locus of two lines, l1 and l2, from $\mathcal{L}_N$. Two situations are then possible.

(i) One or both lines cut p at its edges. The affected edges will then be made of collinear segments from the same line. This is not possible by construction of E: E is built from EP, and no segment formed of collinear segments is accepted in EP (Figure 3.4b). Therefore, l1 and l2 cannot cut p at any of its edges.

(ii) l1 and l2 cut p only at its vertices. We will see that we can then form a new polygon passing through v_i, but with a total number of edges lower than p.

First, we note that neither l1, nor l2 will pass through v_i. This would be in contradiction with the fact that our BFS search produces polygons whose adjacent edges always form a consecutive angle (Figure 3.5b).

Now, let us consider that p has n vertices that we number starting at v_i and sort clockwise: $v_iv_{i+1}v_{i+2}...v_{i+j}...v_{i+n-2}v_{i+n-1}$. l1 will cut p at nonadjacent vertices, v_{i+k} and v_{i+k+l} ($l>1$, $k+l<n$) and l2 at vertices v_{i+r} and v_{i+r+s} ($s>1$, $r+s<n$). $k\neq r$, otherwise A would

not be interior to p. In the following, we will assume k<r, but the same result is obtained assuming the opposite, r<k. Each of the two lines will cut p at a vertex comprised between the vertices of the other line (this follows from Theorem 3.43b, Lee (Lee, 2012)); consequently, r<k+l<r+s.

Let q, be the polygon defined by the clockwise ordered list of vertices (Figure 3.8) $Av_{i+r+s}...v_{i+n-1}v_iv_{i+1}...v_{i+k}$. The edges of q are: the two edges involving A ($\overline{\mathbf{Av_{i+r+s}}}$ and $\overline{\mathbf{v_{i+k}A}}$), then n-1-r-s edges from $v_{i+r+s}...v_{i+n-1}$, the k edges from $v_i...v_{i+k}$ (all of them belong to EP, because their vertices all belong to VP), and the closing edge involving v_i ($\overline{\mathbf{v_{i+n-1}v_i}}$). The total number of edges, m, is equal to: n+k+2-r-s. Because k+l<r+s and l>1, we have that m<n. Therefore, q would have less edges than p and, consequently, c_q (the cycle corresponding to q) would have less edges than c_p. This is in contradiction with the fact that c_p is minimal.

Following the same reasoning we can easily generalize the proof to an arbitrary number of interior points. □

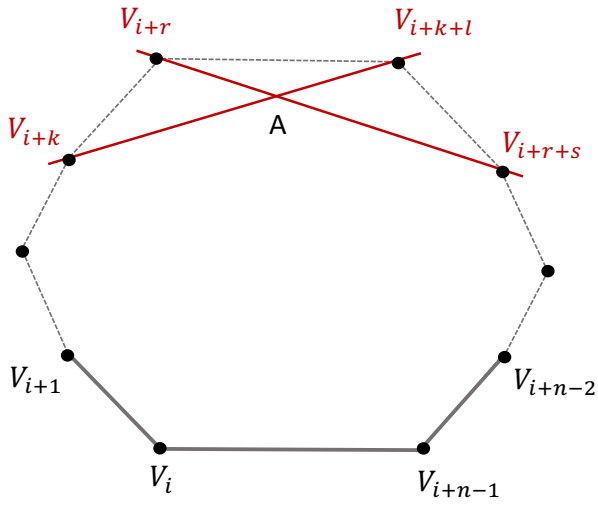


Figure 3.8. Polygons p and q mentioned in the demonstration of Lemma 3. *p is constituted by the more external edges (grey, both continuous and dashed) and q by the list of vertices $Av_{i+r+s}...v_{i+n-1}v_iv_{i+1}...v_{i+k}$; it includes edges from p and two new edges, $\overline{Av_{i+r+s}}$ and $\overline{v_{i+k}A}$, that share point A .*

We can now prove Proposition 4.

Proposition 4. Let c_{SC}, the Shortest Cycle passing through vertices v_i, v_k and v_j in G(V, E); and let p_{SC}, the polygon corresponding to this cycle. Let $p \in \mathcal{P}_N$ be a polygon that passes through v_i and has the two adjacent edges, $\overline{v_iv_j}$ and $\overline{v_iv_k}$. Then, $p_{SC} = p$.

Proof. We know that p_{SC} is a polygon, because c_{SC} is a cycle that has non-repeating edges, which are noncollinear (because in the connectivity matrix, if three arbitrary E vertices, a, b, and c, are such that a*b*c, there will be connections only for the two edges $\overline{ab}$ and $\overline{bc}$, but not for $\overline{ac}$).

We know that p_{SC} is not concave, because a concave polygon does not correspond to a minimal cycle in G. Indeed, if p_{SC} were concave it would have a vertex with an associated angle larger than 180 degrees. The lines intersecting at this vertex would cut p_{SC}, giving children polygons with a lower

number of edges; one of these would contain v_i, having $\overline{v_i v_j}$ and $\overline{v_i v_k}$ as edges. This polygon would have less edges than p_{SC}, which is in contradiction with the fact that p_{SC} is minimal.

From the previous considerations, we know that p_{SC} must be convex. In addition, p_{SC} has no interior points (Lemma 3) and, by construction, all its edges belong to EP. By Proposition 3, there is only a unique convex polygon that satisfies these conditions, therefore $p_{SC} = p$. □

For each vertex v_i we will use BFS to find the shortest cycles corresponding to all the $\overline{v_i v_j}$, $\overline{v_i v_k}$ pairs forming consecutive angles. This procedure will be repeated for all the vertices in G(V, E), guaranteeing, through the use of counters (see conditions C3 and C4 above), that the number of shortest cycles matches that of expected $\mathcal{P}_N$ polygons. By Proposition 4 and conditions C5 and C6 above, we know that the shortest cycles found are unique and will correspond to the $\mathcal{P}_N$ polygons.

3.3.5. Application of the rc_{bd} models to a set of sixteen representative *in silico* tools

In this section, we will use the two versions of the cost framework embodied in equations (7) and (13), respectively, to look at pathogenicity predictors taking into account medical context, compositional properties of the sequenced region, and performance of the tools. In particular, we will see how using the integrative cost framework changes our view on pathogenicity predictors in the context of clinical applications. Theses aspects will be illustrated using a set of sixteen selected pathogenicity predictors (Table 3.1).

Table 3.1. The sixteen pathogenicity predictors used in this work, with their corresponding performance parameters: sensitivity, specificity and coverage.

Predictor	Sensitivity (%)	Specificity (%)	Coverage (%)
FATHMM	83.5	65.8	90.2
LRT	90.2	69.3	86.1
MutationAssessor	86.1	71.9	87.7
MutationTaster	97.9	60.3	99.6
PolyPhen2-HDIV	92.6	63.8	90.9
PolyPhen2-HVAR	87.9	75	90.9
PROVEAN	85.8	79	87.3
SIFT	92.4	68.2	86.6
CADD	99.5	25.4	100
MetaLR	88.7	85.3	99.6
MetaSVM	89.6	87.9	99.6
REVEL	94.2	88.1	99.6
PON-P2	97.4	87.5	49.5
VEST	97.1	82.4	93.7
Pmut	84.7	86.5	90.4
MutPred	95	70.6	28.1

3.3.5.a. Model with no coverage, equation (7)

We apply the rc_{bd} model in (7) when we exclude situation (iii) from our analyses. This is equivalent to either disregard the consequence of the lack of predictions by *in silico* tools, or to implicitly assume that these tools have a 100% coverage. This is common practice in bioinformatics, where *in silico* tools are compared using one or several of the myriad of performance measures based on the confusion matrix (Baldi *et al.*, 2000; Vihinen, 2012b), e.g. MCC, sensitivity, specificity, accuracy, etc.

In Figure 3.9a we show how the application of (7) to the chosen predictors results in sixteen rc_{bd} lines, one per predictor, that cross each other. When, within a sub-interval of $\mathcal{I}$, a line is above the others, this means that the corresponding method has the highest rc_{bd}. For example, this is the case for CADD (Figure 3.9a, light blue line), which is above the other methods for most

of $\mathcal{I}$, indicating that this method is associated to higher budget deviations for the corresponding scenarios. On the contrary, PON-P2 (dark magenta) has the lowest rc_{bd} across much of $\mathcal{I}$, indicating its cost-effectiveness for the corresponding scenarios.

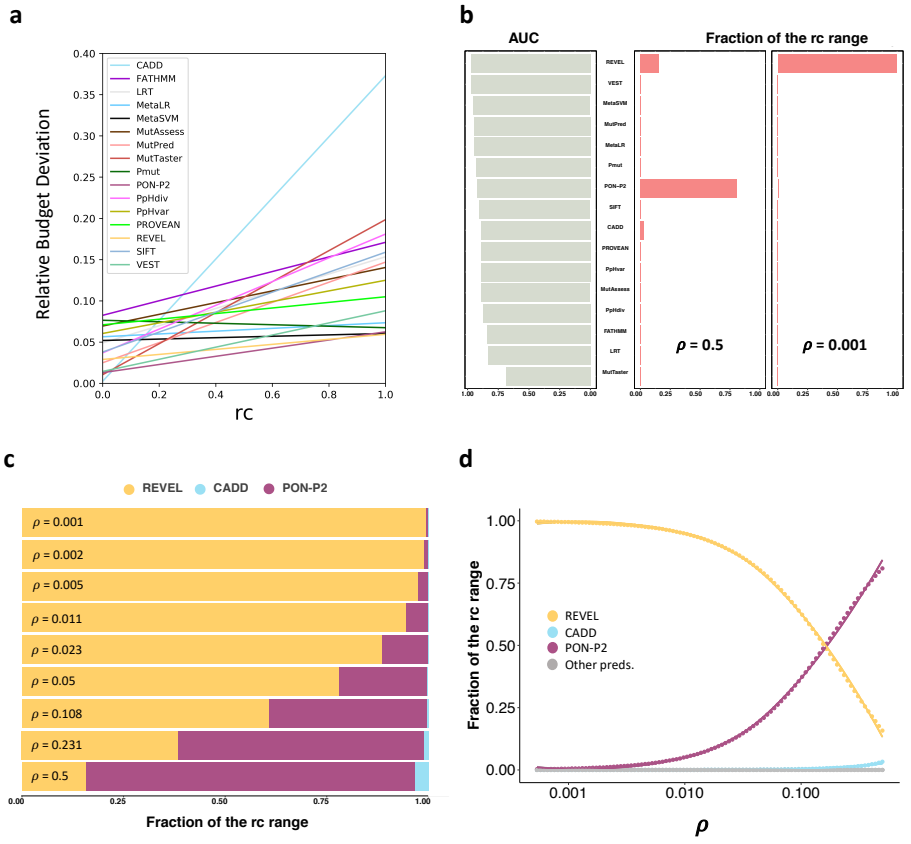


Figure 3.9. Application of the cost framework without coverage (equation (7)) to sixteen pathogenicity predictors and effect of ρ, the fraction of pathogenic variants in the sample, on the distribution of pathogenicity predictors over the cost domain. a) *Line chart representing the relative budget deviation of the sixteen pathogenicity predictors through the rc range.* ***b)*** *Comparison between the fraction of the rc range (fraction of the cost region assigned to each predictor, divided by the total range of the cost region) associated with every predictor and AUC; the analyses were done for ρ=0.5 and ρ=0.001. We see that fraction of the rc range and AUC give very different views on the pathogenicity predictors, indicating how clinical-context considerations may change our view of the applicability of in silico predictors.* ***c)*** *The different divisions of*

the rc range obtained varying ρ between 0.001 and 0.5. The ρ values were chosen so that each order of magnitude is evenly sampled. We see that the predominating predictor, in cost terms, varies drastically, depending on the fraction of pathogenic variants identified in the sequencing experiment. ***d)*** *We illustrate the same phenomenon in more detail, representing the fraction of the rc range occupied by each predictor vs. ρ.*

To quantify the previous analysis, we assign to each method the fraction of $\mathcal{I}$ where its rc_{bd} value is lower than that of the remaining methods, that is, where this method is preferable to the others in terms of cost/budget deviations. In Figure 3.9b we show these fractions for the different predictors, for two ρ values, 0.5 and 0.001. We see that not all the methods are represented, indicating that there are no cost scenarios for which the absent methods are preferred over the others. In fact, only a few predictors participate in the optimal division, in terms of cost, of $\mathcal{I}$: three for ρ=0.5 (Revel, PON-P2 and CADD) and two for ρ=0.001 (Revel and PON-P2). In the figure we also display, for comparison, the AUC values of the sixteen predictors: we see no relationship between these two views of pathogenicity predictors.

In Figure 3.9b we show the results of the cost framework for two ρ values (0.5 and 0.001). We see that each of these values gives a different division of $\mathcal{I}$ in terms of pathogenicity predictors. We further explore this aspect in Figures 3.9c and 3.9d, where we can see how Revel, the leading method at low ρ's, is gradually replaced by PON-P2 as ρ increases.

3.3.5.b. Model with coverage included, equation (13)

Here, we apply the rc_{bd} model in (13), which includes misclassification errors and incomplete coverage, to find the distribution of the sixteen predictors over $\mathcal{T}'$, the set of all possible cost scenarios.

Following the procedure described in the Results section 3.3.4, we build the lines for all possible pair comparisons between methods (Figure 3.10b). These lines divide $\mathcal{T}'$ into a complex mosaic of polygons that once processed gives the desired distribution of methods (Figure 3.10c). We see that five out of the sixteen tools span $\mathcal{T}'$: CADD, MutationTaster, PON-P2, REVEL and VEST. However, $\mathcal{T}'$ is not evenly split among these methods: REVEL covers the major part of it, 77.3%, while the rest only cover small percentages, below 9%. This result is further illustrated in Figure 3.10d, where we also provide the AUC values of each method, for comparison: we see no relationship between these two views of pathogenicity predictors. A method may have the best AUC but it may not necessarily give the minimal budget deviations for a given cost scenario. For example, REVEL has the best AUC (=0.97) and gives minimal rc_{bd} values for an important fraction of $\mathcal{T}'$; however, for cost scenarios for which $rc_0+rc_1\approx1$ (near the hypotenuse of $\mathcal{T}'$) or when rc_1 tends to zero, PON-P2 (hypotenuse region of $\mathcal{T}'$; AUC=0.92) and CADD (bottom region of $\mathcal{T}'$; AUC=0.89), respectively, are preferable. More extreme examples can be found, like the case of MetaSVM, which has the third highest AUC (=0.95), but is not preferred in any cost scenario. This contrast between AUC and rc_{bd} views of predictors, already observed in the previous section, reflects the fact that rc_{bd} integrates different aspects of the prediction problem, from the performance of the predictor to the cost of its application.

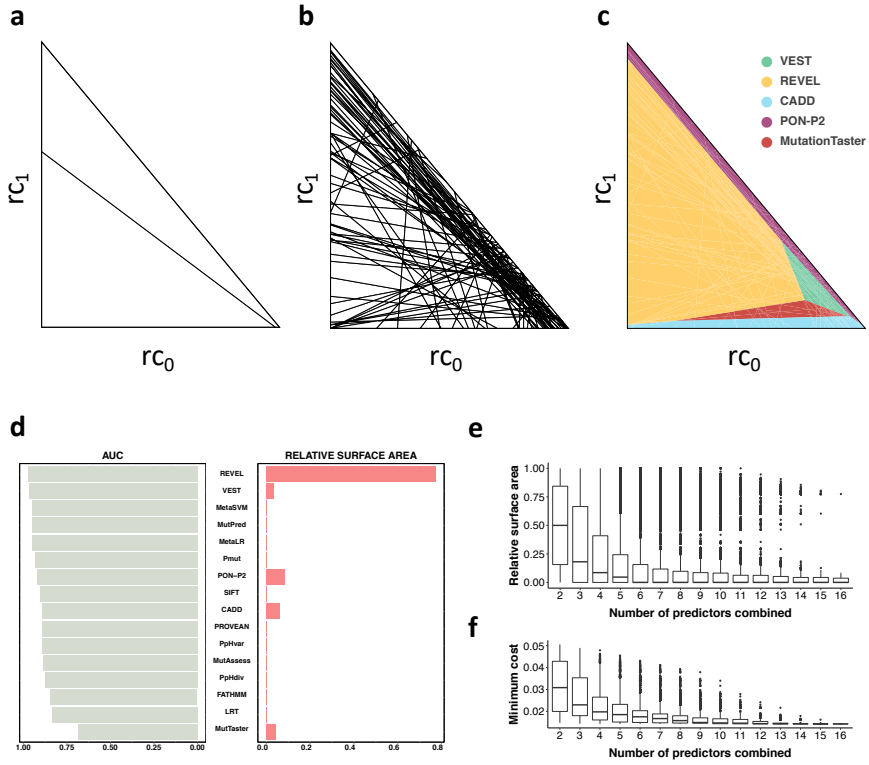


Figure 3.10. Application of the cost framework with coverage (equation (13)) to sixteen pathogenicity predictors. a) *The triangle corresponds to the cost space, of in silico tools. When comparing two predictors in terms of application cost, the line crossing the triangle separates the scenarios where each method has a lower average cost.* ***b)*** *As we add more predictors, we obtain an intricate pattern of polygons; inside each polygon, one predictor is preferred over the others, in terms of cost.* ***c)*** *Unification of the polygons corresponding to each predictor using our approach (Results section 3.3.1) results in regions (one different color per method) allowing the identification of the optimal predictor for every cost scenario.* ***d)*** *Comparison between the relative surface area of the cost domain (surface area of the cost region assigned to each method, divided by the total area of the triangle) associated with every predictor and AUC. The differences between both measures illustrate how clinical-context considerations may change our view of the applicability of in silico predictors.* ***e)*** *Distributions of the relative surface areas for the different combinations of predictors (the X-axis is the number of predictors combined). As we combine more predictors, a preferred set of methods emerges; the remaining methods tend to occupy smaller sections of the cost domain.* ***f)*** *Same as* ***e)****, but for the minimum* rc_{bd} *(Y-axis) instead of*

relative surface areas. We see that the more methods we consider, the more economic becomes the resulting division. All the analyses were done for ρ=0.5.

Of course, other combinations of methods can be used to cover $\mathcal{T}$ and our procedure can find their best distribution in the cost space. This is important when we want to replace a method because it is either no longer available to the community or supported by its creators. In Figure 3.10e, we see that different combinations of methods give broadly varying divisions of $\mathcal{T}$, and even for two methods the possible situations may go from equipartition to no-partition (one method is enough to cover $\mathcal{T}$). However, it is also clear that not all divisions produce equally low rc_{bd} (Figure 3.10f) and that our procedure may find the combination of methods giving the best partition.

Study of the impact of different ρ values (Figures 3.11a-3.11b) completes the description of our approach. We find that the overall picture may not change so much, because only two to five methods out of sixteen, participate in the final division of $\mathcal{T}$. However, their identities change, showing that when choosing pathogenicity predictors, we cannot ignore the expected fraction of pathogenic variants in the sample. This may be important for users of *in silico* tools that are working with variants from gene regions with specific conservation patterns and/or from patient samples with concrete ethnic origins. As before, the relationship between the vie of predictors provided by AUC and that of the cost framework are in sharp contrast (Figure 3.11c).

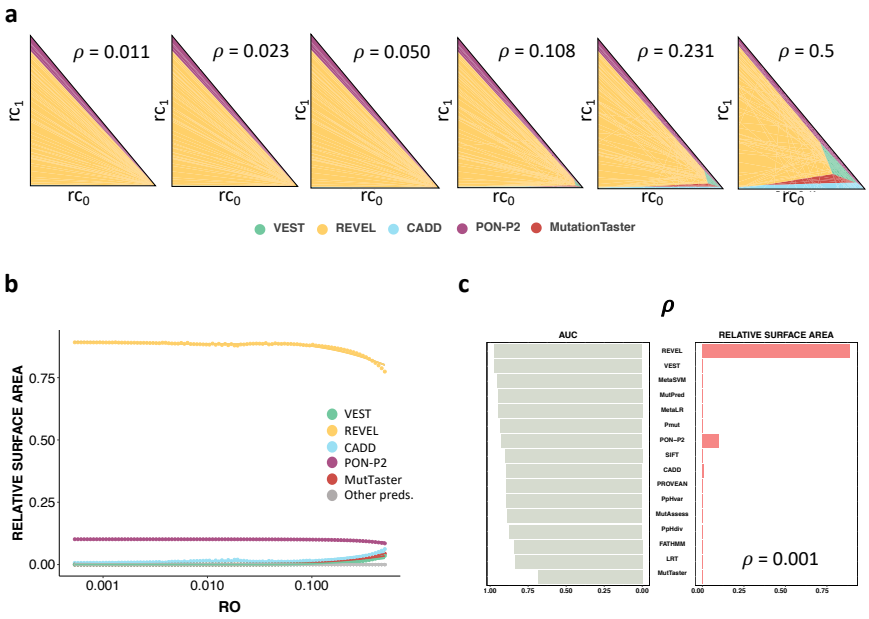


Figure 3.11. Effect of ρ, the fraction of pathogenic variants in the sample, on the distribution of pathogenicity predictors over the cost domain. a) *The different divisions of the cost domain triangle obtained varying ρ between 0.011 and 0.5. The ρ values were chosen so that each order of magnitude is evenly sampled. We see that the predominating predictor, in cost terms, does not vary substantially, depending on the fraction of pathogenic variants identified in the sequencing experiment. However, we can observe a certain variability for the rest of the predictors.* ***b)*** *We illustrate the same phenomenon in more detail, representing the relative surface area occupied by each predictor vs. ρ.* ***c)*** *Comparison between the relative surface area of the cost domain (surface area of the cost region assigned to each predictor, divided by the total area of the triangle) associated to every predictor and AUC. This figure is equivalent to Figure 3.10d, here obtained for ρ=0.001. The conclusion is the same: the differences between both measures illustrate how clinical-context considerations may change our view of the applicability of in silico predictors.*

3.4. Discussion

In this work we present a framework for assessing and comparing the clinical applicability of pathogenicity predictors, inspired on the use of cost models for clinical tests (Pepe, 2003). We have extended the conventional formalism (Adams and Hand, 1999; Drummond and Holte, 2006; Hernández-

Orallo, Flach and Ferri, 2012) derived for binary classifiers to predictors with incomplete coverage; that is, those tools that do not always produce an answer to the prediction problem. This situation is relatively frequent in the case of pathogenicity predictors (Vihinen, 2012b; de la Campa, Padilla and de la Cruz, 2017) and has consequences, in the clinical context, that cannot be ignored.

We have applied this framework to a set of sixteen selected predictors, unveiling a view of these tools completely different from that obtained when using only standard performance parameters (Figures 3.9b, 3.10d and 3.11c). Indeed, if we employ AUC or MCC, the differences between predictors are gradual; it is difficult to decide which one is preferable among the top performers (Figure 3.12). However, when using our model the situation changes dramatically. First, we see that only one method is optimal for most of the cost scenarios. The identity of the predominant method changes depending on a property of the sequenced region, ρ, the frequency of pathogenic variants. For high and low values of ρ, PON-P2 and REVEL prevail, respectively, when coverage is ignored in the cost framework (Figures 3.9c-3.9d; Equation (7)). When introducing coverage (Equation (13)), REVEL prevails over the whole range of ρ values (Figure 3.11b), although other methods are also represented, e.g. PON-P2, VEST and CADD. These methods do not necessarily have top MCC or AUC values (Figure 3.12), but their combination of sensitivity, specificity and coverage makes them optimal for certain cost scenarios.

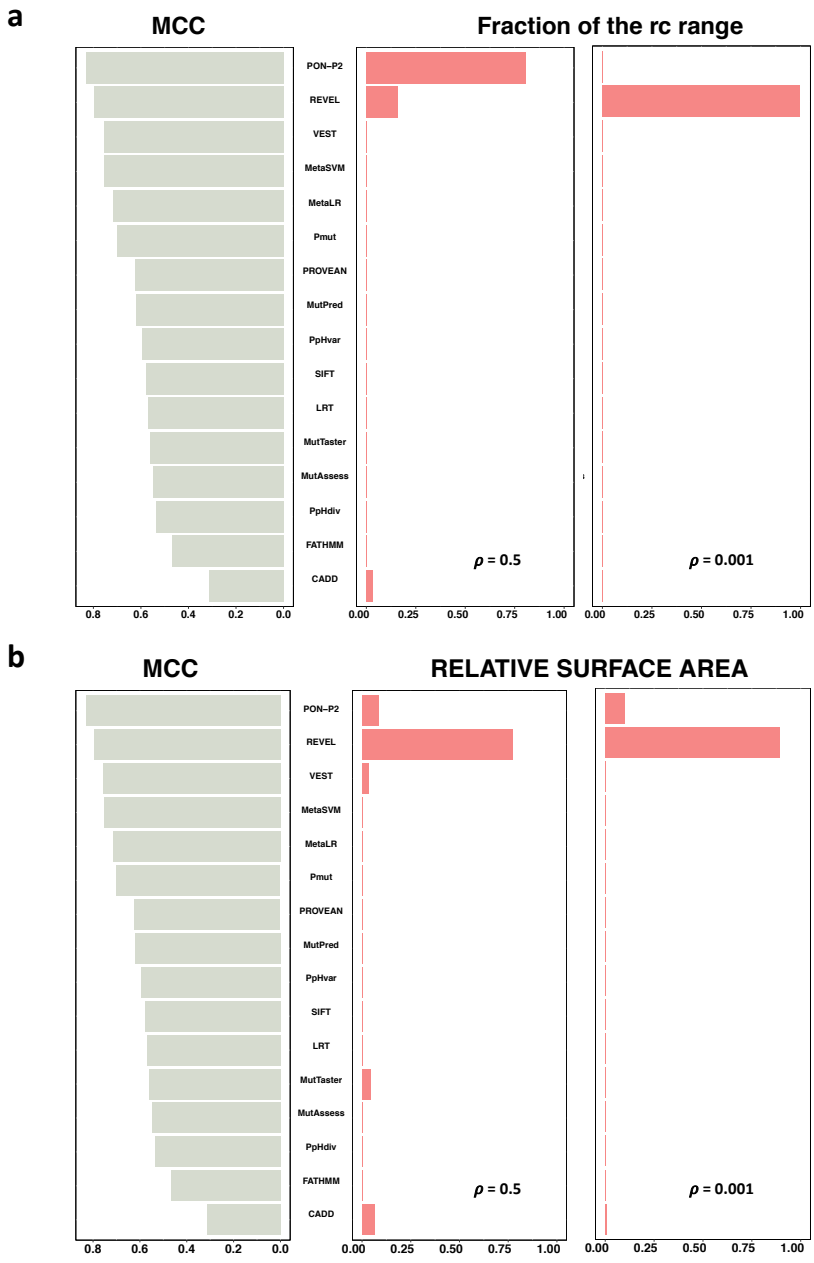


Figure 3.12. Comparison between different measures to estimate a pathogenicity predictor performance. a) *Comparison between the fraction of the rc range associated with every predictor and MCC.* ***b)*** *Comparison between the relative surface area of the cost domain (surface area of the cost region assigned to each predictor, divided by the total area of the triangle) associated with every predictor and MCC. Figures* ***a)*** *and* ***b)*** *are equivalent to Figures 3.9b, and Figures 3.10d and 3.11c, respectively, obtained for*

MCC instead of AUC. The conclusion is the same: the differences between both measures illustrate how clinical-context considerations may change our view of the applicability of in silico predictors.

The previous results support the idea that, when choosing *in silico* tools for clinical applications, it is as important to consider performance as clinical scenario (medical consequences/costs). This would be less relevant if sequencing costs and clinical scenarios were always the same; however, this is not the case. First, the costs of sequencing change between countries (Schwarze *et al.*, 2018). And second, in terms of impact we cannot dissociate the result of the NGS experiment from the downstream medical decisions and consequences, whose costs may vary substantially, depending on factors such as budget and drug price differences between countries (Barbieri *et al.*, 2005; Smith, Busse and Schreyo, 2008; *Health at a Glance 2019*, 2019; Czech *et al.*, 2020), within countries (Care, Services and Medicine, 2013), etc. In this situation, healthcare centers may need to apply different strategies to budget their sequencing studies, taking into account the cost of misclassification errors, the population involved in the study, etc. All these factors are captured by the parameters rc_0, rc_1 and rc_2 in our cost framework, which can then be used to identify the pathogenicity predictor most adequate for the needs of the planning center. It has to be mentioned, however, that the exact values of these parameters are difficult to estimate; it is easier for professionals to estimate their ratios (Adams and Hand, 1999). For example, saying that the cost of missing a patient (reflected in rc_0) is larger than twice the cost of treating a healthy individual (reflected to rc_1). Our cost framework then allows to incorporate this type of analysis in a simple way: the ratios between rc_0 and rc_1 values correspond to lines passing through the origin (Figure 3.13). They allow to partition the cost scenarios and identify which predictors are preferable within them. Also, when costs associated to coverage are small

compared to those of misclassification error, we can revert to the model with no coverage included.

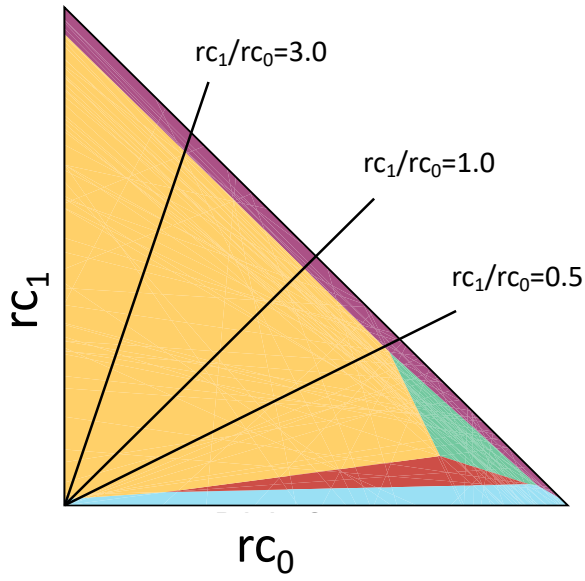


Figure 3.13. Incorporation of the cost ratios to the cost framework. *The cost ratios between rc_0 and rc_1 values correspond to lines passing through the origin of the triangle. From top to bottom, we represent three different clinical scenarios: (i) the cost of treating a healthy individual (rc_1) is three times bigger than the cost of missing a patient (rc_0), (ii) the costs of treating a healthy individual and missing a patient are equal, and (iii) the cost of missing a patient is twice the cost of treating a healthy individual.*

In summary, our cost framework provides a tool for the analysis of pathogenicity predictors in terms of the impact they will have in downstream medical decisions. This will allow users in healthcare settings to identify what tool is more adapted to their specific clinical context.

3.5. Conclusions

In this work we present a framework for assessing and comparing the clinical applicability of pathogenicity predictors, extending the conventional

formalism (Adams and Hand, 1999; Drummond and Holte, 2006; Hernández-Orallo, Flach and Ferri, 2012) to adapt it to the cases of incomplete coverage, which in the clinical context has consequences that cannot be ignored. Our results unveil a view of pathogenicity predictors completely different from that obtained when using only standard performance parameters. Using standard performance parameters makes it difficult to decide which pathogenicity predictor is preferable among the top performers (Figure 3.12). However, this situation changes dramatically when using our model. We see that although some methods do not have top MCC or AUC values, their combination of sensitivity, specificity and coverage makes them optimal for certain cost scenarios. Moreover, depending on the frequency of pathogenic variants (ρ), we observe that there are variations of the predominant method for the different cost scenarios. Summing up, our results show that when choosing *in silico* tools for clinical applications, it is important considering both performance as well as clinical scenario (medical consequences/costs), and our cost framework provides a tool that will allow users in healthcare settings to identify what predictors are more adapted to their specific clinical context.

4. THE RELATIONSHIP BETWEEN MOLECULAR IMPACT AND DISEASE PHENOTYPE IN THE CONTEXT OF CPDs

The results presented in this chapter have been published in Scientific Reports (Marín*, Ò., Aguirre*, J., and de la Curz, X. (2019). Compensated pathogenic variants in coagulation factors VIII and IX present complex mapping between molecular impact and hemophilia severity. *Scientific Reports, 9, 9538.* *https://doi.org/10.1038/s41598-019-45916-3*). *Equally contributing authors.

The goal of this chapter is to study the relationship between molecular impact and disease severity in hemophilia cases in the context of compensated pathogenic deviations (CPDs), studying data from FVIII and FIX coagulation factors. Moreover, as there is growing amount of evidence showing that genetic background may contribute to clinical phenotype, we will study the genetic background of hemostasis proteins.

4.1. Introduction

As we have seen in previous chapters, understanding the phenotypic consequences of genetic variability is still an open challenge relevant to different areas of biology, from biomedical research (Knight, 2009; Riera, Lois and de la Cruz, 2014) to protein evolution studies (DePristo, Weinreich and Hartl, 2005; de Visser and Krug, 2014; Zhang and Yang, 2015; Storz, 2016). A case of particular interest is that of the human sequence variants known as compensated pathogenic deviations (CPDs) (Kondrashov, Sunyaev and Kondrashov, 2002), which are damaging for human carriers but appear as neutral in other species (Figure 4.1a). This dual aspect of the amino acid replacement reflects the two main characteristics of CPDs. First, in its human protein location, the amino acid replacement has an impact on protein structure/function big enough to cause disease. Second, in the non-human protein, this impact is modulated by a suppressor mechanism. Kondrashov et al. (Kondrashov, Sunyaev and Kondrashov, 2002) identified compensatory mutations as the main suppressor mechanism (the so-called the Compensatory Hypothesis (Xu and Zhang, 2014)) and postulated that such mutations most likely correspond to substitutions at spatial locations near CPDs (Figure 4.1b). The compensatory hypothesis is strongly supported by a series of studies involving large structural analyses (Barešić *et al.*, 2010),

stability computations (Xu and Zhang, 2014), and comparative genomics (Jordan *et al.*, 2015).

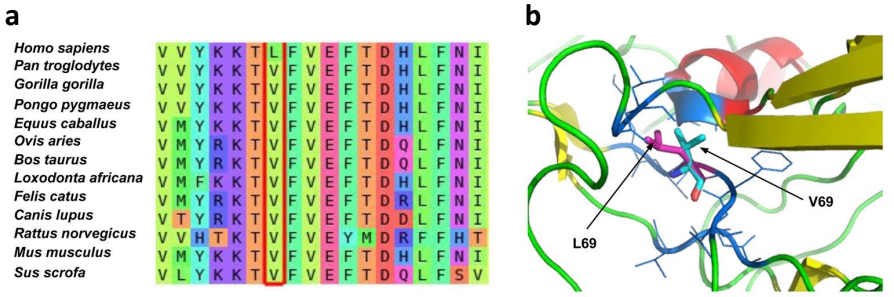


Figure 4.1. Compensated pathogenic deviations (CPDs): definition. a) *Illustration of the concept of CPDs: a variant that is pathogenic in humans but neutral in other species. The variant shown in the figure, L69V, affects human FVIII and leads to hemophilia A* (Xue *et al.*, 2010). *The location of this variant in the multiple sequence alignment of the FVIII family (in the red box) shows that valine appears to be native in other species, such as chimpanzees, mice, and rats.* ***b)*** *Spatial neighborhood (dark blue residues) of the CPDs (native and mutant residues indicated by light blue and magenta sticks, respectively), where compensatory mutations are more likely to happen* (Kondrashov, Sunyaev and Kondrashov, 2002).

Within this stream of research, Ferrer-Costa et al. (Ferrer-Costa, Orozco and de la Cruz, 2007) explored whether the molecular nature (e.g., protein location, changes in biophysical properties) of pathogenic deviations (PDs) determines the probability of compensation. They found that CPDs are usually less structurally disruptive than average PDs, as they are associated with higher solvent exposure and smaller changes in physico-chemical properties. This result was confirmed by Barešić et al. (Barešić *et al.*, 2010), who also found that CPDs tend to avoid residues directly involved in protein function (e.g., from binding and catalytic sites).

Together, these studies provide a good understanding of CPDs at the molecular level. However, the relationship between molecular and organismal phenotypes, which is key to the evolutionary study of these variants, remains

largely unexplored. In general, biomedical evidence suggests that there is a monotonic correspondence between the molecular impact of PDs and the clinical phenotype. For example, studies examining the development of pathogenicity predictors (Ferrer-Costa, Orozco and de la Cruz, 2002; Riera, Lois and de la Cruz, 2014) have consistently observed that small changes in molecular properties typically correspond to variants that have no impact on the individual's health. Other studies identified the relationship between the range of the mutation effect and clinical severity. For example, Miyata et al. (Miyata, Miyazawa and Yasunaga, 1979) represented the impact of variants with a continuous function of distance in the physico-chemical space. They related the values of this function to the severity of hemolytic anemia and found correspondence between the two phenotype levels. For G6PD deficiencies, Miller and Kumar (Miller and Kumar, 2001) found a similar trend, again using physico-chemical differences to measure the impact of a mutation. In addition, a comparable relationship between molecular properties and clinical severity was observed when the effects of a mutation were represented by experimental $\Delta\Delta G$ (protein stability change upon mutation) (Randles *et al.*, 2006) or conservation-based measures (related to functional role of the mutated residue) (Botstein and Risch, 2003).

Considering that CPDs are a subset of PDs (Ferrer-Costa, Orozco and de la Cruz, 2007; Barešić *et al.*, 2010), we expect to find a similar correspondence between molecular impact and clinical phenotype (i.e., "mild-to-mild" for most CPDs and "severe-to-severe" for the small fraction of structurally/functionally disruptive CPDs). Confirmation of this relationship would allow us to study CPDs using known models of protein evolution, which utilize measures of molecular impact as a proxy for fitness (DePristo, Weinreich and Hartl, 2005). However, the situation may be more complex, as a growing amount of evidence shows that genetic background may contribute to clinical phenotype (Badano and Katsanis, 2002; Storz, 2016), particularly

those genes belonging to the functional module (or disease pathway) of the variant carrier protein (Zaghloul and Katsanis, 2010). For example, in the case of hemophilia, it is known (Pavlova and Oldenburg, 2013) that genetic alterations in hemostasis proteins mitigate the clinical symptoms of the disease, and several reports relate disease severity and the effect of genetic background (Tsoutsman, Bagnall and Semsarian, 2008; Kleffmann *et al.*, 2012).

In this chapter, we characterize the relationship between molecular impact and organismal phenotype in the context of CPDs. As a model system, we use the coagulation factors VIII (FVIII) and IX (FIX), two proteins whose variants cause hemophilia A and B, respectively. Hemophilia is a well-known disease –it was first reported in Jewish writings dating the 2nd century AD (Ingram, 1976)—that primarily affects males. It has a characteristic bleeding phenotype (Srivastava *et al.*, 2013), the severity of which affects organismal fitness to different degrees (i.e., bleeding can be either mild or life-threatening). In the context of the present chapter, it is important to note two interesting aspects of research on hemophilia. First, information about disease severity is available for a large number of variants (see section 4.2, Materials and Methods) of FVIII and FIX, which will allow us to analyze the correspondence between different measures of molecular impact (at the structure and function levels) and organismal fitness, using severity as a proxy for the latter. Second, the functional module of FVIII and FIX, constituted by the proteins from the hemostasis system, has been well-described (Stassen, Arnout and Deckmyn, 2004; Pavlova and Oldenburg, 2013; Versteeg *et al.*, 2013; Ribeiro *et al.*, 2015). Thus, we can assess its mutational load in the general population (based on the 1000 Genomes Project (Auton *et al.*, 2015)), which is relevant for understanding the modulatory potential of genetic background.

4.2. Materials and Methods

4.2.1. CPD dataset

To obtain our sets of CPDs for FVIII and FIX, we followed a three-step protocol. First, for both coagulation factors, missense pathogenic variants were retrieved from the databases CHAMP for Hemophilia A (http://www.cdc.gov/ncbddd/hemophilia/champs.html) (Payne *et al.*, 2013) and CHBMP for Hemophilia B (http://www.cdc.gov/ncbddd/hemophilia/chbmps.html) (Li *et al.*, 2013). The pathogenicity of these variants was validated using ClinVar (Landrum *et al.*, 2016). We identified two and three cases for FVIII and FIX, respectively, that we considered benign according to ClinVar (Landrum *et al.*, 2016). These were removed from our dataset. Thus, we obtained a total of 971 variants for FVIII and 391 for FIX. We annotated these variants with the severity phenotype provided, which in this case was the bleeding patterns. Second, we built a multiple sequence alignment (MSA) for the two coagulation factors, as described below. Third, for each variant, we checked if in the MSA location of the wild-type residue we could find the mutant residue in another species. When this was the case, the variant was considered a CPD. At the end of this process, we obtained 122 (87 mild, 35 severe) and 47 (25 mild, 22 severe) CPDs for FVIII and FIX, respectively. In some analyses, we used noCPDs. We obtained 849 (406 mild, 443 severe) and 344 (93 mild, 251 severe) noCPDs for FVIII and FIX, respectively. The complete list of mild/severe variants used in this chapter/work is provided in Appendix 1 Additional File 9.1.1.

In the Results section 4.3.3 “The molecular impact of CPDs in FVIII…” we compare the molecular properties of CPDs ($\Delta\Delta G$, Blosum62 matrix elements and Shannon’s Entropy) associated to mild and severe versions of the disease. For this comparison we had to partition the set of CPDs into two subsets, corresponding to the variants associated to mild and severe versions

of the disease, respectively. The size of these subsets was relatively small (87 mild and 35 severe for FVIII; 25 mild and 22 severe for FIX) making the comparisons more sensitive to experimental error (Kanyongo *et al.*, 2007), which in our case corresponds to the uncertainty level in the causality assignment of the variants. To reduce this effect to a minimum, we manually verified the causality annotations of each CPD, using the references provided in the CHAMP/CHBMP databases. In particular, we checked to which extent the criteria employed to establish causality were comparable to the most recent recommendations in the field (MacArthur *et al.*, 2014): we looked for evidence such as (Bell *et al.*, 2011; MacArthur *et al.*, 2014) uniqueness of the variant in the carrier's sequence, use of healthy individuals as controls, and structural/functional analysis of the variant's impact and conservation at the mutation locus. We discarded those CPDs for which the evidence of causality was unclear (that is, it was either not mentioned or appeared to be weak). At the end of this process, the final number of CPDs was: 91 (62/29 corresponding to mild/severe disease) for FVIII and 25 (12/13 corresponding to mild/severe) for FIX. Given the small sample size of the FIX dataset, we decided to limit the comparisons in Figure 4.4 to FVIII and present the results for FIX in the Appendix 1 Figure 9.1.1. The final sets of manually curated CPDs are provided in Appendix 1 Table 9.1.1.

For the remaining proteins, CPDs were obtained as follows. First, we queried the UniProt (UniProt-Consortium, 2014) database with the keywords "lethal/severe" and "mild". Of the resulting set of proteins, we kept only those for which there were at least five instances of each case. We then followed the second and third steps of the protocol for FVIII and FIX (described in the previous paragraph): we constructed an MSA for each protein and examined the MSA columns of the human native residues to determine whether there were pathogenic residues in the non-human species. At the end of this process, we had retrieved 155 mild and 229 severe variants that led to 17 mild

and 25 severe CPDs distributed over 14 proteins (Figure 4.4 And Appendix 1 Additional File 9.1.1). In some analyses, we used noCPDs, of which there were 138 and 204 mild and severe cases, respectively.

The final list of variants is provided in Appendix 1 Additional File 9.1.1.

4.2.2. Characterization of variants in terms of molecular properties

In this study, the molecular impact of variants is described using four parameters: protein stability change upon mutation ($\Delta\Delta G$), solvent accessibility, elements of the BLOSUM62 matrix, and Shannon's entropy at the mutation locus. These parameters, or related ones, are routinely used to characterize pathogenic variants and reflect different aspects of their impact on protein structure and function (Riera, Lois and de la Cruz, 2014). In particular, $\Delta\Delta G$ is a central parameter in the biophysical theory of protein evolution (DePristo, Weinreich and Hartl, 2005; Stenson *et al.*, 2012), and it was recently used by Xu and Zhang (Xu and Zhang, 2014) to test the compensation hypothesis. We estimated $\Delta\Delta G$ using the FoldX suite (van Durme *et al.*, 2011). Relative solvent accessibility (obtained from the experimental structures of FVIII—PDB code 2R7E—and FIX—PDB codes 1CFH, 1IXA and 3LC5), which indicates whether a variant may be structurally disruptive or affect protein-protein interactions (Ferrer-Costa, Orozco and de la Cruz, 2002), was computed with the NACCESS program (Hubbard and Thornton, 1993). BLOSUM62 matrix elements, obtained by Henikoff and Henikoff (Henikoff and Henikoff, 1992) from the frequency of amino acid exchanges in blocks of aligned sequences from conserved protein regions, capture some aspects of molecular evolution (Pearson, 2013). It has been shown (Rudnicki, Mroczek and Cudek, 2014) that BLOSUM matrices summarize the changes in physico-chemical properties (hydrophobicity, size, charge) associated with amino acid substitutions and related to changes in

protein function and structure. This parameter was employed, among others, by Ferrer-Costa et al. (Ferrer-Costa, Orozco and de la Cruz, 2007) to show that CPDs are milder than PDs. Finally, Shannon's entropy at the mutation locus in the MSA is a measure of the conservation patterns at this position in the MSA of the protein family (Valdar, 2002), which is related to functional/structural restraints. It is equal to: $-\sum_i p_i \,.\, \log(p_i)$, where *i* runs over all the amino acids at the MSA column corresponding to the mutated residue. Shannon's entropy varies between 0 and 4.322, with low and high values indicating highly and poorly conserved locations, respectively.

4.2.3. Multiple sequence alignments

For each protein in our dataset, we built a corresponding MSA by (i) retrieving from Ensembl (Yates *et al.*, 2016) the mammalian orthologs of the human protein and (ii) aligning them with the program Muscle (Edgar, 2004).

4.2.4. Hemostasis proteins

The primary biological roles of FVIII and FIX are to contribute to hemostasis (Versteeg *et al.*, 2013). This defensive mechanism is responsible for minimizing the blood loss resulting from vascular injury through the coordinated action of several proteins (Versteeg *et al.*, 2013). Both biomedical/clinical (Pavlova and Oldenburg, 2013) and evolutionary (Ribeiro *et al.*, 2015) studies have shown that variants in these proteins can modify the bleeding patterns of their carriers, as key phenotype of hemophilia. On this basis, for the genetic background of FVIII and FIX, we used a list of 19 proteins (17 cases plus FVIII and FIX; Appendix 1 Table 9.1.2) that was recently compiled (Ribeiro *et al.*, 2015) to study the genomic basis of phenotypic variation in hemostasis.

4.2.5. Variants in the 1000 Genomes Project

In section 4.3.4 "Genetic variability in hemostasis proteins" of this chapter, we examined the amount and composition of the sequence variants present in hemostasis proteins in a population of healthy individuals. We obtained the relevant data from the 1000 Genomes Project (Auton *et al.*, 2015). Specifically, we retrieved all the missense variants in the 19 hemostasis proteins listed in Appendix 1 Table 9.1.2 carried by each male individual in the 1000 Genomes Project database.

4.3. Results

4.3.1. CPDs in FVIII and FIX can be associated with either mild or severe forms of hemophilia

For the two coagulation factors, we found that both their CPDs and non-compensated pathogenic deviations (noCPDs) are associated to either mild or severe forms of hemophilia (Figure 4.2a). The percentages are specific for each protein: for CPDs 29% (FVIII) and 47% (FIX) of the cases are associated to severe disease; for noCPDs these figures rise to 52% (FVIII) and 73% (FIX). For both coagulation factors, CPDs are less frequently associated with severe symptoms than noCPDs (Fisher's exact test: p-value=1.0×10^{-6} for FVIII and p-value=4.0×10^{-4} for FIX).

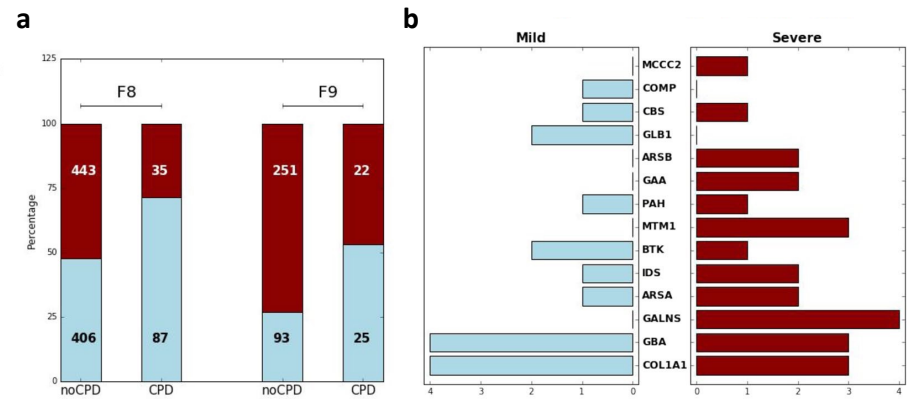


Figure 4.2. Compensated pathogenic deviations (CPDs): distribution relative to disease severity. a) *Distribution of mild (blue) and severe (red) cases associated with CPDs and noCPDs variants for FVIII and FIX.* ***b)*** *Distribution of mild and severe CPDs in genes for which disease severity annotations are available.*

We investigated other diseases in which CPDs are spread over the severity scale. To do so, we used severity annotations and variant information retrieved from the UniProt database. The number of cases was small, comprising 42 CPDs (17 associated with mild disease and 25 associated with severe disease) distributed over 14 genes (Figure 4.2b). For this reason, we could not draw statistically relevant conclusions for each gene. However, we found that CPDs may be associated with either mild or severe forms of disease (Figure 4.2b). Treating the whole dataset as a single sample revealed no detectable differences between CPDs and noCPDs (Fisher's exact test: p-value=1). This result does not contradict the trends observed for FVIII and FIX since pooling data from different diseases may obscure gene-specific trends (Riera, Padilla and de la Cruz, 2016).

4.3.2. CPDs in FVIII and FIX tend to be mild at the molecular level

Next, we characterized the molecular impact of FVIII and FIX CPDs to determine whether they tend to be milder than noCPDs, as found in the

general case (Ferrer-Costa, Orozco and de la Cruz, 2007; Barešić *et al.*, 2010). To this end, we compared the distribution of CPDs and noCPDs for a series of properties that reflect complementary aspects of molecular impact: change in free energy upon mutation (ΔΔG, used in biophysical models of protein evolution and here computed using FoldX (van Durme *et al.*, 2011)), solvent accessibility at the mutation locus in the experimental structure (a measure of the potential for structure disruption of mutations), elements of the BLOSUM62 matrix (which capture evolutionary information (Pearson, 2013) and can be related to the physico-chemical changes associated with amino acid replacement (Rudnicki, Mroczek and Cudek, 2014), and conservation pattern (measured using Shannon's entropy, which is related to the functional and structural role of the native residue (Botstein and Risch, 2003)) at the mutation locus in the multiple sequence alignments of the FVIII and FIX families.

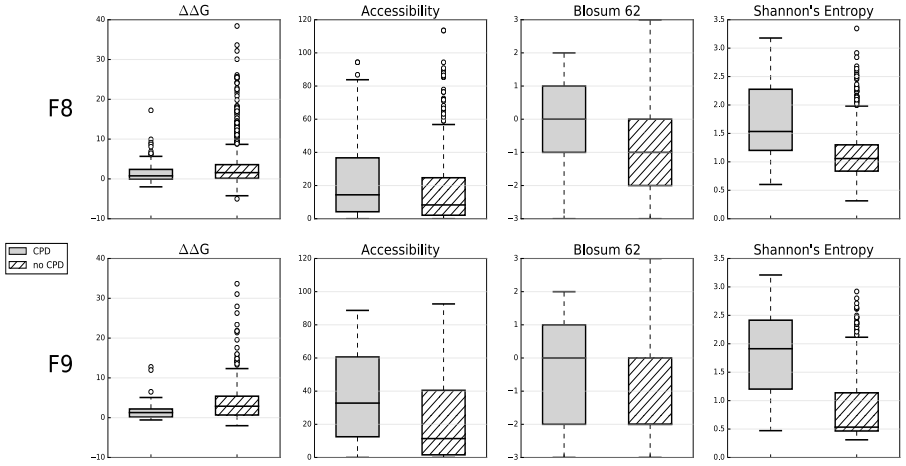


Figure 4.3. Differences between CPDs and noCPDs. *For variants of the two coagulation factors in our dataset, FVIII (top) and FIX (bottom), we computed the values of four properties: ΔΔG (change in protein stability upon mutation), relative solvent accessibility, BLOSUM62 elements and Shannon's entropy. Using boxplots, we then separately represented the value distributions for the CPDs (grey) and noCPD (striped) variants. There is a statistically significant tendency for noCPDs to adopt slightly more*

extreme values than CPDs, indicating that the latter are molecularly "milder" than the former.

We observed the same situation for both coagulation factors (Figure 4.3): a significant trend for CPDs to be less disruptive than noCPDs. The p-values for the Mood's median test for FVIII are the same (Table 4.1), p-value=0, for all the properties ($\Delta\Delta$G, relative solvent accessibility, BLOSUM62 elements and Shannon's entropy). The corresponding values for FIX are as follows (Table 4.1): 0 for $\Delta\Delta$G, relative solvent accessibility and Shannon's entropy, and $4.1x10^{-14}$ for BLOSUM62 elements. In spite of the significant differences we observed, there is an overlap between the distribution of CPDs and noCPDs in all cases, indicating that some CPDs may be as molecularly disruptive as some noCPDs. For example, if we consider $\Delta\Delta$G values in the case of FVIII, the median of the CPDs distribution is above the $\Delta\Delta$G value of 63% of no CPDs (Figure 4.3). That is, from the perspective of free energy change upon mutation, 50% of CPDs are more disruptive than 37% of noCPDs. For FIX, the situation is similar, with 50% of CPDs being more disruptive than 29% of noCPDs.

Table 4.1. Summary of the results of the statistical tests corresponding to the comparisons shown in the different figures.

Figure	Test	p-value
Figure 4.2A, FVIII	Fisher's exact	1.0x10^-6
Figure 4.2A, FIX	Fisher's exact	4.0x10^-4
Figure 4.2B	Fisher's exact	1
Figure 4.3, FVIII, ΔΔG	Mood's median	0
Figure 4.3, FVIII, Acces.	Mood's median	0
Figure 4.3, FVIII, Bl62	Mood's median	0
Figure 4.3, FVIII, Shan. Entr.	Mood's median	0
Figure 4.3, FIX, ΔΔG	Mood's median	0
Figure 4.3, FIX, Acces.	Mood's median	0
Figure 4.3, FIX, Bl62	Mood's median	4.1x10^-14
Figure 4.3, FIX, Shan. Entr.	Mood's median	0
Figure 4.4, FVIII, ΔΔG	Mood's median	0
Figure 4.4, FVIII, Bl62	Mood's median	2.8x10^-4
Figure 4.4, FVIII, Shan. Entr.	Mood's median	0.01
Appendix 1 Figure 9.1.1, FIX, ΔΔG	Mood's median	0.54
Appendix 1 Figure 9.1.1, FIX, Bl62	Mood's median	0.94
Appendix 1 Figure 9.1.1, FIX, Shan. Entr.	Mood's median	0.60

4.3.3. The molecular impact of CPDs in FVIII (and FIX) is not strongly related to disease severity

The overlap between the distributions of CPDs and noCPDs in Figure 4.3 suggests that CPDs associated to severe forms of disease (Figure 4.2A) could correspond to highly disruptive mutations. To determine the extent to which this was true, we explored whether our data support a correspondence between molecular impact and disease severity. To this end, we split the CPD populations into two groups: those leading to mild and severe forms of hemophilia. We then compared these two groups in terms of molecular-level properties examined before.

Splitting the original CPD datasets involves a reduction of the initial sample, making any ensuing comparison more sensitive to causality assignment errors (see section 4.2.1). To minimize this effect, we worked with

a subset of the original CPD datasets with high-quality causality annotations (see section 4.2.1). For FVII, the CPD sample went from 122 to 91 cases and for FIX it went from 47 to 25 cases. Then, for the comparisons in this section, these datasets were partitioned into two groups: (i) CPDs associated to mild disease and (ii) CPDs associated to severe disease. In the case of FVIII, the corresponding groups had 62 and 29 CPDs, respectively, and in the case of FIX, they had 12 and 13 CPDs, respectively.

Comparison of FVIII CPDs leading to mild and severe disease (Figure 4.4) produces statistically significant results for all properties (Mood's median test, Table 4.1): $\Delta\Delta G$ (p-value=0), Shannon's entropy (p-value=0.001), and BLOSUM62 elements (p-value=2.8×10^{-4}). However, visual inspection of the results (Figure 4.4) shows different degrees of overlap between the mild and severe distributions, consistent with deviations from a mild-to-mild/severe-to-severe relationship between molecular impact and severity phenotype. The result for $\Delta\Delta G$, for which the distribution overlap is moderate, suggest that the relationship may be valid for extreme values of $\Delta\Delta G$.

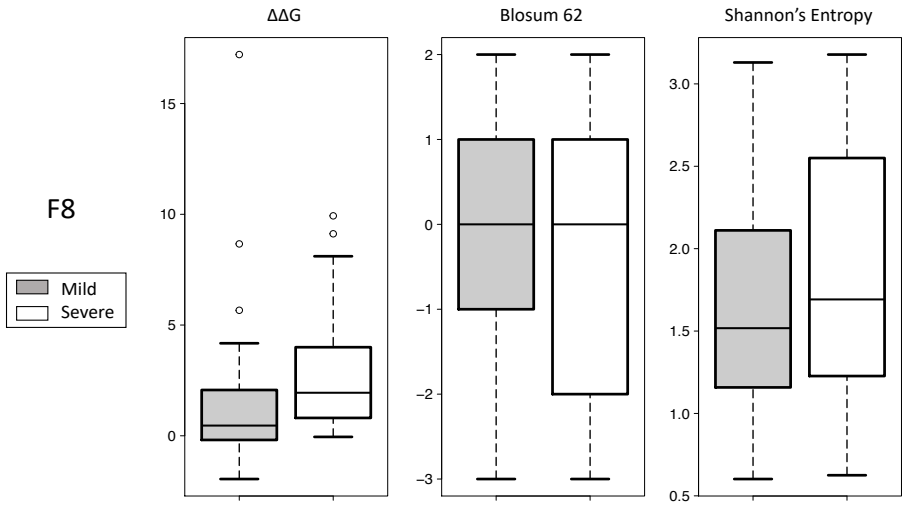


Figure 4.4. The molecular impact of CPDs and clinical severity. *For FVII, we plotted the value distribution of three properties (ΔΔG, BLOSUM62 matrix elements and Shannon's entropy) for the severe (white) and mild (grey) subsets.*

For Shannon's entropy, the difference between medians is surprising from a functional point of view of conservation, because the distribution for severe phenotypes is shifted towards non-conserved locations. However, the difference is small (<0.25), particularly when we consider the substantial overlap between entropy distributions (Figure 4.4). In our case, the difference between medians may reflect aspects specific to the compensation of highly disruptive variants. These variants, frequent among the CPDs associated to severe disease (ΔΔG plot in Figure 4.4), are usually harder to compensate (Ferrer-Costa, Orozco and de la Cruz, 2007). For this reason, we expect to find them in 3D environments where sequence changes are numerous and provide better chances of compensation (Barešić *et al.* 2010). In these environments, the loci of both the CPD and its neighbors will have larger entropies, and this may be reflected in the median shift described.

For FIX (Appendix 1 Figure 9.1.1) the trends are similar to those of FVIII, with median differences in the same directions and large overlaps between distributions. In this case, none of the comparisons were statistically

significant (Table 4.1) suggesting, together with the visual analysis, the presence of deviations from the monotonic relationship between molecular impact and severity. These results must be considered with care given the small sample size for FIX.

4.3.4. Genetic variability in hemostasis proteins

In parallel with the previous analyses, we characterized the inter-individual variability in hemostasis proteins because evidence from biomedical (Pavlova and Oldenburg, 2013) studies shows that genetic alterations in these proteins can modulate the bleeding phenotype of hemophilia. To this end, we mapped the variants carried by 1233 males (obtained from the 1000 Genomes Project (Auton *et al.*, 2015)) to a set of known hemostasis proteins (Ribeiro *et al.*, 2015) (19 proteins, including FVIII and FIX, Appendix 1 Additional File 9.1.1). We then analyzed the resulting data in terms of amount and nature (i.e., pathogenic or neutral) of missense variants, two measures of the genetic alterations of the disease pathway related to disease severity (Muntoni *et al.*, 2006; Tsoutsman, Bagnall and Semsarian, 2008; Kelly and Semsarian, 2009; Bauce *et al.*, 2010; Bergmann *et al.*, 2011; Kleffmann *et al.*, 2012).

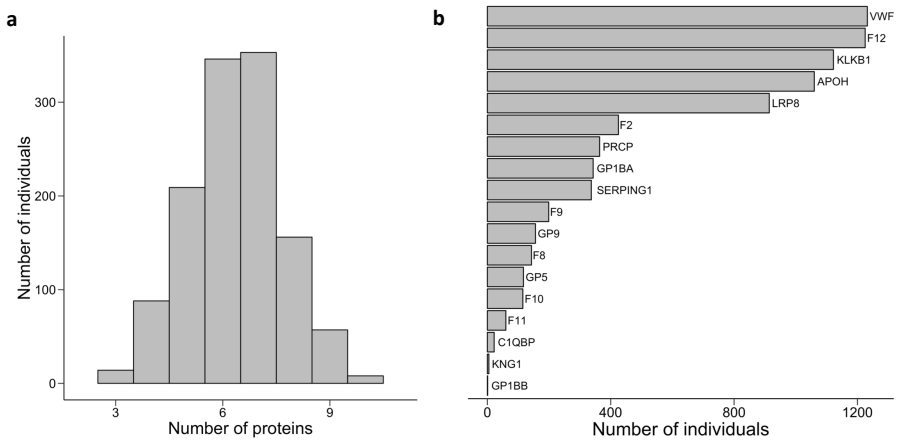


Figure 4.5. Variability of the genetic background of FVIII and FIX. *This figure shows concrete aspects of how the sequence variants from the 1000 Genomes Project are distributed across hemostasis proteins.* ***a)*** *Frequency histogram of the number of mutated proteins per individual.* ***b)*** *Number of individuals for which each hemostasis protein appears to be mutated.*

As Figure 4.5a shows, all individuals in the population present variants in at least three of the proteins, and most frequently (in 699 of 1233), individuals had variants in 6-7 proteins. However, not all the proteins are equally mutated; the von Willebrand factor (vWF) and coagulation factor FXII (F12) were mutated in almost all individuals (Figure 4.5b), while variants in Kininogen-1 (KNG1) and Platelet glycoprotein 1b beta chain (GP1BB) were seldom observed.

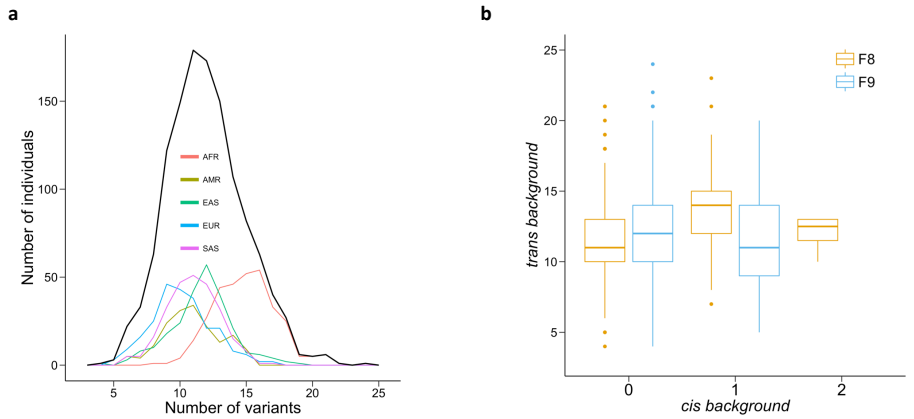


Figure 4.6. Variability in the genetic background of FVIII and FIX. *This figure shows concrete aspects of how the sequence variants from the 1000 Genomes Project are distributed across hemostasis proteins (i.e., the genetic background of FVIII and FIX).* ***a)*** *Frequency histogram of the number of variants per individual. The results for the whole population are shown in black, and separate results for each of the five super-populations in the 1000 Genomes Project—African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS)—are shown in color.* ***b)*** *Distribution of the number of variants of background proteins relative to FVIII (yellow) and FIX (blue).*

The number of variants changes between individuals (Figure 4.6a), mainly ranging from 5-20, and is affected by ethnicity. Plotting the number of variants in hemostasis proteins (excluding those of FVIII and FIX) relative to those in FVIII and FIX (Figure 4.6b) indicates different possibilities for variability in genetic background. This variability, following the notation in Jordan et al. (Jordan *et al.*, 2015) (*cis:* in the same protein, *trans:* in a different protein), sometimes may be completely *trans* relative to either FVIII or FIX, and sometimes it may be a mixture of *cis* and *trans* variants. The latter is relevant because *cis* locations are believed to host compensatory variants (Kondrashov, Sunyaev and Kondrashov, 2002) more frequently than *trans* locations.

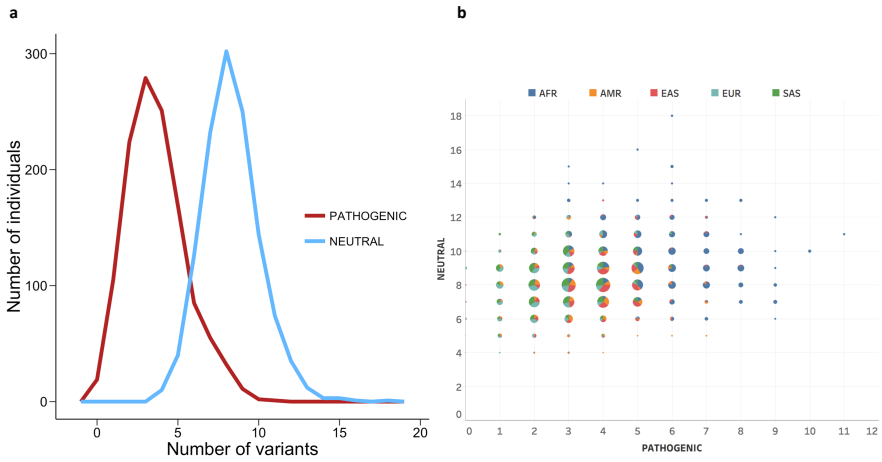


Figure 4.7. Pathogenic load of hemostasis proteins. a) *Distribution of the number of neutral (blue) and pathogenic (red) variants per individual in the 1000 Genomes population.* ***b)*** *Scatterplot showing the different combinations of neutral and pathogenic variants found in the population. The size of circles represents the number of individuals in which each combination was observed. In addition, each circle is a pie plot that represents the fraction of individuals from the different superpopulations in the 1000 Genomes Project: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS).*

At the compositional level, we distinguished between neutral and pathogenic variants and looked for inter-individual differences in the number of each. For all the identified variants, we queried the HGMD (Stenson *et al.*, 2012) database to retrieve all available pathogenicity annotations. Figure 4.7a shows that pathogenic variants are present in a majority of the population (98%), although neutral variants predominate. We refine this view in Figure 4.7b, which shows that both variant types appear in different combinations and that some individuals have higher number of pathogenic variants in hemostasis proteins than others.

4.4. Discussion

Biophysical studies of CPDs using structural analyses and stability computations have explained their dual (i.e., pathogenic/neutral) behavior

(Miyata, Miyazawa and Yasunaga, 1979; DePristo, Weinreich and Hartl, 2005; Ferrer-Costa, Orozco and de la Cruz, 2007; Xu and Zhang, 2014; Jordan *et al.*, 2015). In particular, we know that compensatory mutations are the principal mechanism suppressing the harmful effects of CPDs (Xu and Zhang, 2014; Jordan *et al.*, 2015) and that these effects, in terms of molecular properties, tend to be milder than those of non-compensated PDs (Ferrer-Costa, Orozco and de la Cruz, 2007; Barešić *et al.*, 2010) (Figure 4.3). The biophysical approach has been extended to explain the appearance of CPDs during evolution using $\Delta\Delta G$ as a proxy for fitness either explicitly or implicitly (DePristo, Weinreich and Hartl, 2005; Ferrer-Costa, Orozco and de la Cruz, 2007; Barešić *et al.*, 2010; Sikosek and Chan, 2014). More precisely, dePristo et al. (DePristo, Weinreich and Hartl, 2005) proposed a formalism in which, upon mutation, fitness variations are expressed as an exponential function of the difference in $\Delta\Delta G$ from a reference value. This formalism is easily interpretable, and the authors illustrated its potential in the comparison of two competing hypotheses about the origin of CPDs. However, the explanatory power of the model is limited to those cases where there is a monotonic (mild-to-mild/severe-to-severe) correspondence between molecular impact and organismal fitness, and the effect of genetic background is small.

In our work, we explored the extent to which this is the case for CPDs in FVIII and FIX. Specifically, we studied (i) how measures of molecular impact (structural and functional) relate to the severity phenotype (Figures 4.3, 4.4 and Appendix 1 Figure 9.1.1) and (ii) the compositional properties of genetic background (Figures 4.5, 4.6, 4.7). For FVIII, $\Delta\Delta G$ was the molecular property showing the most noticeable difference between the mild and severe distributions (Figure 4.4). For FIX, the trend is comparable (Appendix 1 Figure 9.1.1) but statistically non-significant. This result suggests that, at least for

FVIII, fitness models based on ΔΔG (DePristo, Weinreich and Hartl, 2005; Echave and Wilke, 2017) may be useful for evolutionary study of CPDs. However, the applicability range of these models may be restricted when working with ΔΔG estimates because of their moderate correlation with observed stability changes. For example, in the case of FoldX (Guerois, Nielsen and Serrano, 2002) the authors cite a value of 0.8 (r^2=0.64); for the same program, Tian et al. (Tian *et al.*, 2010) find a correlation of 0.5 and a low accuracy (69.5%) for the discrimination between stabilizing and destabilizing variants. On the other hand, results from the application of FoldX to the characterization of mutations causing Fabry disease indicate (Riera *et al.*, 2015) that extreme ΔΔG values may successfully identify pathogenic variants. On this basis, we believe that, when working with computational estimates of ΔΔG, it may be preferable to restrict the use of ΔΔG-based fitness models to those CPDs with a large effect on stability.

For CPDs having a small effect on stability, the applicability of ΔΔG-based models is more limited. This may occur for two reasons, apart from the previously discussed problems that arise when working with ΔΔG estimates. The first reason is a low correlation between ΔΔG and protein function (Sánchez *et al.*, 2006; Rost and Bromberg, 2009); a CPD may have a small impact on ΔΔG but large impact on protein function, resulting in a noticeable effect on fitness. However, models only based on ΔΔG would predict a minor effect on fitness. The second reason may be the modulatory effect of genetic background. Evidence from both experimental and theoretical bioinformatics studies shows that the phenotypic effect of mutations is modulated by genetic background (Breen *et al.*, 2012; Rockah-Shmuel, Tóth-Petróczy and Tawfik, 2015; Vu *et al.*, 2015; Hou *et al.*, 2016; Storz, 2016). For example, by performing RNAi experiments with two *C. elegans* isolates, Vu et al. (Vu *et al.*, 2015) found that about 20% of the ~1400 genes they tested displayed

background-dependent differences in the severity of the loss-of-function phenotype. An array of biomedical studies also supports the regulatory role of background (Muntoni *et al.*, 2006; Tsoutsman, Bagnall and Semsarian, 2008; Kelly and Semsarian, 2009; Bauce *et al.*, 2010; Bergmann *et al.*, 2011; Kleffmann *et al.*, 2012). For example, To-Figueras et al. (To-Figueras *et al.*, 2011) found that in congenital erythropoietic porphyria, a disease caused by mutations in *UROS* (an enzyme of the erythroid heme biosynthesis pathway), severity depends on the variants present in *ALAS2*, the rate-controlling enzyme of this pathway. In the case of hemophilia, we know that genetic background must be considered because specific variants in hemostasis proteins other than FVIII and FIX can modify severity phenotype (Pavlova and Oldenburg, 2013). Within this context, one expects that the cumulative effect of background variants on fitness may sometimes surpass that of CPDs with a small effect on stability, thus limiting the applicability of $\Delta\Delta$G-based models in this case.

In the previously cited biomedical studies, 1-3 (usually pathogenic) variants in the genes of the disease pathway are enough to modulate the effect of the causal variant. In our case, after characterizing the number and kinds of variants in hemostasis proteins, we found that many individuals already carry 5-20 variants (Figure 4.6a) and, frequently, one or more of these variants are pathogenic (Figure 4.7, section 4.2). Another interesting aspect of our results (Figure 4.6, 4.7) is the diversity they reveal; neither the background size nor composition are constant in the population (due to ethnic diversity and inter-individual variability). Comparable results are observed at the variant level; the same variant may appear with different backgrounds in different individuals (Figure 4.8).

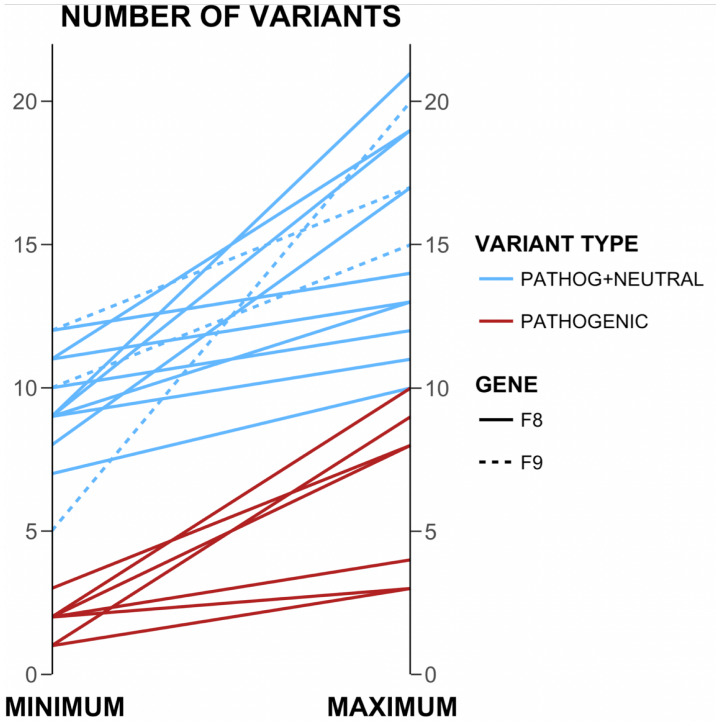


Figure 4.8. Differences in the background of specific variants between FVIII and FIX. *The genetic background of a given variant can vary between individuals. Here, we focus on variants of FVIII and FIX and define background as the number of accompanying variants in the hemostasis proteins using population data from the 1000 Genomes Project. Each line represents a variant in these coagulation factors (continuous and broken lines for FVIII and FIX, respectively) that is present in more than one individual. The line unites the minimum (left axis) and maximum (right axis) number of background variants observed for that variant. Blue indicates that all the background variants were counted, regardless of their nature, and red indicates that only pathogenic background variants were counted. For the latter investigation, we found only examples relating to FVIII.*

Given their impact on protein stability (Yue, Li and Moult, 2005) and protein-protein interactions (Zhong *et al.*, 2009), we expect that the coincidence of several pathogenic variants in the same individuals will have a net lowering effect on the efficacy of the hemostasis mechanism. This effect will change between individuals since the number of pathogenic variants varies between individuals (Figure 4.7) and because the net effect of the

variants may not follow a simple additive model (Wells, 1990). In summary, the genetic background of FVIII and FIX has the potential to modulate the impact of CPDs.

4.5. Conclusions

Our results are specific for FVIII and FIX; for these coagulation factors, they suggest that, from an evolutionary point of view, we need to expand our models for the appearance of CPDs during evolution. Present models (DePristo, Weinreich and Hartl, 2005) may work only for the most disruptive variants (involving large $\Delta\Delta G$ values or affecting functional sites); including the contribution of genetic background (e.g., applying the approach proposed for complex epistatic effects (Reva, Antipin and Sander, 2011; de Visser and Krug, 2014) under a biophysical framework (Sikosek and Chan, 2014)) may be relevant when the variants under study have a mild impact on protein function. Extension of this conclusion to other proteins will require additional work showing, among other things, that molecular factors related to protein function (e.g., protein interactions, complexity of the functional module, etc.) support a modulatory role for genetic background; however, they must be accompanied by an effort to increase the accuracy of $\Delta\Delta G$ computations. Otherwise, increasing the complexity of the models will augment the error of fitness estimates.

5. STUDY OF *IN SILICO* PREDICTORS IN A PRIMARY IMMUNODEFICIENCY (PID) GENE PANEL

The goal of this chapter is to study the application of *in silico* predictors to the characterization of the variants identified with gene sequencing panels, using as a model the panel designed for the diagnosis of patients of Primary Immunodeficiency Disease (PID), developed in the Immunology and Autoinflammatory diseases' groups, at the Vall d'Hebron University Hospital.

5.1. Introduction

The introduction of NGS in clinical diagnosis (Bertier, Cambon-Thomsen and Joly, 2018; Di Resta *et al.*, 2018) has led to an increasing number of diseases and disorders – cardiovascular diseases, immunodeficiency diseases, etc. – for which multigene panel tests are available (BlueShield, 2020). In fact, several studies point in the direction of using gene panels as an alternative to WGS/WES as a first-line test for well-defined phenotypes (Brunelli *et al.*, 2019; Marques Matos, Alonso and Leão, 2019; Sun *et al.*, 2019). This is supported by studies showing that gene panels improve diagnostic yield compared with WES (Y. Xue *et al.*, 2015; Di Resta *et al.*, 2018). For example, the reanalysis of a hearing loss gene panel has resulted in an increase of the diagnostic rate from 39 to 43%, diagnosing nine patients thanks to newly published evidence, adoption of new interpretation guidelines and expanded analysis range (Sun *et al.*, 2019). However, in spite of these advances, the variant interpretation problem is still severe in the case of gene panels and the performance of *in silico* tools in this context remains an open issue.

In this chapter we characterize the performance of *in silico* tools using a Primary immunodeficiency (PID) Gene Panel as a model. PIDs are a heterogeneous group of inherited disorders caused by a variety of monogenic immune defects. Currently, more than 360 genes involved in

immunodeficiencies have been identified and classified by the International Union of Immunological Societies (IUIS) (Picard *et al.*, 2018). Patients suffering from PIDs usually present with a varying degree, unusual/severe infections, autoimmunity, autoinflammation, allergy, etc. A correct clinical diagnosis (identifying the exact type of PID) has crucial consequences in terms of prognosis, treatment and genetic counseling (Yska *et al.*, 2019). However, the phenotypic and genotypic heterogeneity of PIDs, which causes atypical presentations and overlap of symptoms between diseases, generates an unclear genotype-phenotype correlation. The lack of a clear correlation makes genetic diagnosis in patients with PIDs complex and laborious, impeding to reach a definitive molecular diagnosis in many cases (Nijman *et al.*, 2014; Yska *et al.*, 2019). With the application of NGS moving to earlier stages in the diagnostic pipeline for primary immunodeficiencies (PIDs) (Yska *et al.*, 2019), the aim of this study is to explore the performance of *in* silico tools for variant interpretation when used in the context of gene panels, and see how we can improve it. To this end, we focus on the following points: (i) to characterize the genetic diversity captured by the Primary Immunodeficiency Gene Panel, (ii) to characterize the performance of standard pathogenicity predictors in the annotation of the variants identified when using this panel; and (iii) explore approaches to the processing of the gene panel variants that can complement standard methods.

5.2. Materials and Methods

5.2.1. Patient and variant dataset

The starting point of this project is a set of 226 panels corresponding to the same number of patients of Primary Immunodeficiency (PID). These data have been kindly provided by the Immunology and Autoinflammatory diseases' group, at the Vall d'Hebron University Hospital. Variants identified in

these panels were annotated using the VEP annotation software (Yates *et al.*, 2016) release 95, from January 2019. To extract the missense variants, we only retrieved the variants that were tagged as "missense_variant" or "missense_variant,splice_region_variant".

The overall number of variants identified in all the panels was 38626. These variants were subsequently utilized to characterize the behavior of pathogenicity predictors.

5.2.2. Pathogenicity predictors

These tools generate a numerical score that, after comparison with a cutoff value, is utilized to classify the target variant as either neutral or pathogenic. Sometimes, for different reasons (Vihinen, 2020), the predictor does not give a result for a given variant.

In this work, we estimated the performance for a set of fifteen representative pathogenicity predictors: PolyPhen2-HDIV (Adzhubei *et al.*, 2010), PolyPhen2-HVAR (Adzhubei *et al.*, 2010), SIFT (Ng and Henikoff, 2003), CADD (Kircher *et al.*, 2014), MutationTaster (Schwarz *et al.*, 2014a), MutationAssessor (Reva, Antipin and Sander, 2011), REVEL (Ioannidis *et al.*, 2016), FATHMM (Shihab *et al.*, 2013), LRT (Chun and Fay, 2009), PROVEAN (Choi *et al.*, 2012), MetaLR (Dong *et al.*, 2015), MetaSVM (Dong *et al.*, 2015), VEST (Carter *et al.*, 2013), MutPred (Pejaver *et al.*, 2017) and DANN (Quang, Chen and Xie, 2015). For each variant we retrieved the pathogenicity prediction for these tools from the dbNSFP database.

5.2.3. Neutral and pathogenic variants

To estimate the performance parameters of the fifteen pathogenicity predictors, we retrieved a set of 1790 pathogenic and 1111 neutral missense variants from the UniProt/SwissProt (Bateman *et al.*, 2017) database.

5.2.4. Performance assessment and coincidence rules

The variant dataset in section 5.2.3 was utilized to estimate the performance of the different pathogenicity predictors. To this end, we applied each of the chosen predictors to the variant dataset, and then we computed seven standard measures of success rate (Baldi *et al.*, 2000; Vihinen, 2013; Riera *et al.*, 2015; Riera, Padilla and de la Cruz, 2016): sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, Matthews correlation coefficient (MCC) and coverage (α). They were computed as follows:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$$

$$\alpha = \frac{N}{N_{tot}}$$

In all previous equations: TP (true positives) and FN (false negatives) correspond to the number of correctly and incorrectly identified pathogenic variants, respectively. TN (true negatives) and FP (false positives) correspond to the number of correctly and incorrectly identified neutral variants, respectively. N is the total number of annotations generated by the predictor. Finally, N_{tot} is the total number of variants in the dataset.

The coincidence rule utilized to combine predictors is the one advanced in the ACMG-AMP guidelines (Richards *et al.*, 2015): predictions combined from different *in silico* tools are considered as a single piece of evidence, accepting coincident predictions and rejecting noncoincident ones.

5.2.5. Building the panel-specific predictor

We used a Random Forest (RF) to build our panel-specific predictor. To this end, we utilized the RF software available in the Scikit-learn (Pedregosa *et al.*, 2011) package, with default parameters. From the total of 2901 variants in our dataset (see section 5.2.3), 75% (2175) were used to train the RF, and 25% (726) to test it and compare its performance to other pathogenicity predictors. The RF was trained with ten features, six of them related to the protein stability and conservation impact: change in hydrophobicity, change in the amino acid volume, the element of the Blosum62 matrix (Henikoff and Henikoff, 1992), corresponding to the amino acid replacement, Shannon's entropy (Cover and Thomas, 2006), position-specific scoring matrix (Henikoff and Henikoff, 1994), and the functional importance of the residue. The remaining four features were retrieved from the score of four pathogenicity predictors: MetaSVM, MetaLR, DANN and REVEL. The performance estimates

were obtained following a standard five-fold cross-validation procedure, and the final success rate of the RF predictor was measured with the test set.

5.2.6. Variants in the 1000 Genomes Project

In the last section of this chapter, we examined the amount and composition of the sequence variants present in a population of healthy individuals. We obtained the relevant data from the 1000 Genomes Project (Auton *et al.*, 2015). Specifically, we retrieved all the missense variants corresponding to the proteins listed in the gene panel carried by each individual in the 1000 Genomes Project database.

5.2.7. Computations

All the scripting for data management and small data analyses was done using Python 3.6.

To compare the distribution and composition of variants in patients and healthy individuals populations (individuals from 1000 Genomes Project (Auton *et al.*, 2015)), we performed a clustering analysis. The clustering was done using the t-SNE (t-distributed Stochastic Neighbor Embedding) dimensionality reduction algorithm (van der Maaten and Hinton, 2008), as implemented in Scikit-learn (Pedregosa *et al.*, 2011) package.

5.3. Results and Discussion

5.3.1. The genetic diversity captured by the Primary Immunodeficiency (PID) Gene Sequencing Panel

In this section, we analyze the Primary Immunodeficiency (PID) Gene Sequencing Panel in terms of its properties more directly related to genetic diversity, which is the main output of the sequencing process. For the set of

226 patients, the diagnostic yield of this gene panel is about 10%, with 24 out of the 226 patients correctly diagnosed. This result is within the range diagnostic yields of NGS in PID reported in a systematic review by Yska et al. (Yska *et al.*, 2019), who cite values comprised between 15 to 79%. In a previous study, Nijman et al. (Nijman *et al.*, 2014) reported a diagnostic yield of 15%, with three patients describing an atypical presentation of previously described PIDs. This study underlines a relevant fact, which is that the heterogeneity of the disease can be a factor limiting the diagnostic yield of NGS in PIDs. However, technical factors can also contribute to hamper the diagnostic process, starting with the expected number of variants obtained when using the panel. In this sense, it is worth noting that the panel is composed by 323 genes that mostly code for proteins with sizes comprised between 150 and 1000 amino acids (Figure 5.1). Because early results from my Master Thesis shows the existence of a linear relationship between protein size and number of variants (Figure 1.9), we decided to explore this aspect of the panel, to determine whether protein length could be a hidden confusing factor in the diagnosis process, with large proteins contributing more variants than short proteins.

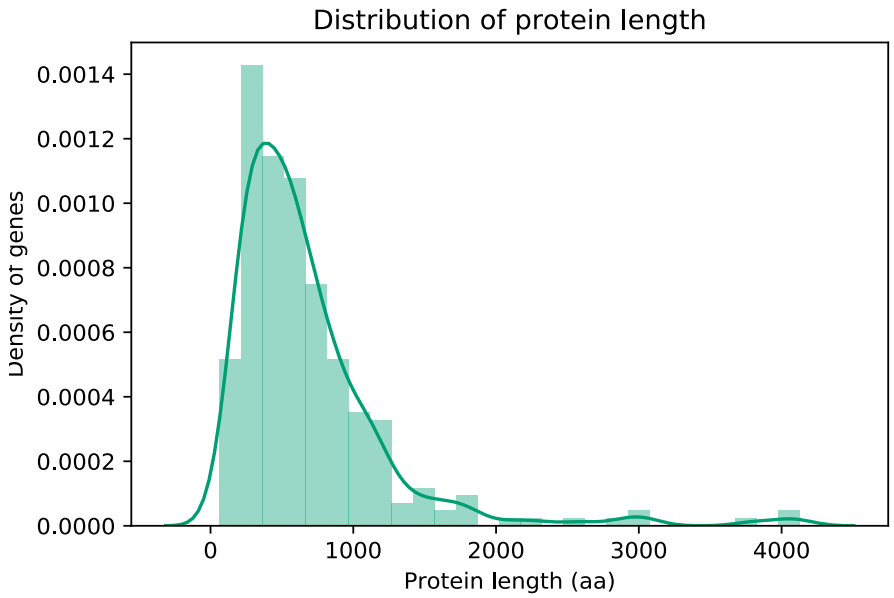


Figure 5.1. Distribution of protein length (aa) of the Primary Immunodeficiency Gene Sequencing Panel.

Thus, we obtained the raw amount of variants identified when using the panel in PID patients. We found that there was an average of 4 missense variants per gene and patient in the panel. This result is consistent with the work of Andrews and colleagues (Andrews, Sjollema and Goodnow, 2013), who found that, on average, 2% of the human population carry a missense mutation in any given gene.

After plotting the number of variants per protein as a function of protein size, grouping data from the different patients, we do not find any clear relationship (Figure 5.2), indicating that protein size is not a confusing factor in the diagnosis process.

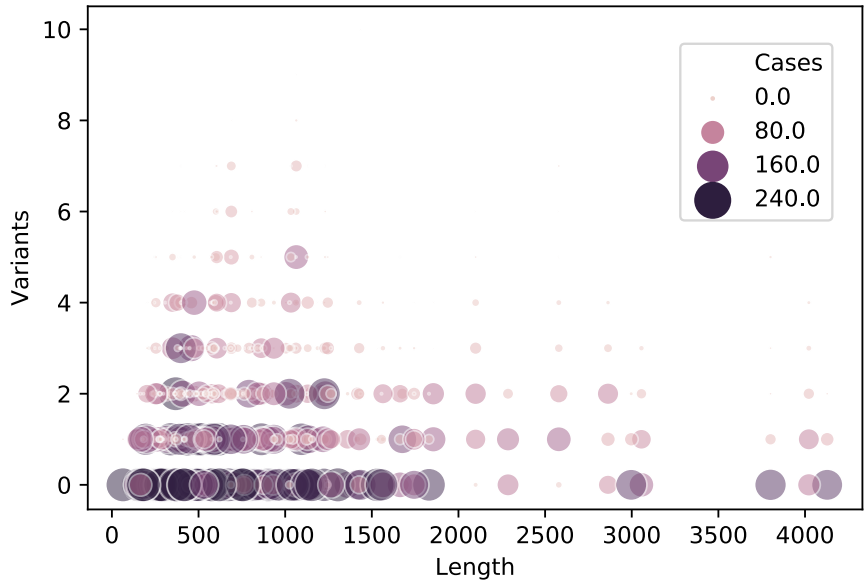


Figure 5.2. Scatter plot of number of missense variants retrieved for each patient. *The circle size represents the number of individuals for which a concrete number of missense variants was retrieved for a specific protein length.*

Use of *in silico* tools to characterize the pathogenic nature of the variants identified by the panel

The use of *in silico* tools for the annotation of sequence variants with medical purposes is mostly regulated by the use of the ACMG-AMP guidelines (Richards *et al.*, 2015). These guidelines state the following: (i) several methods should be utilized to score sequence variants; (ii) the results will be utilized only when there is coincidence in all the predictions. Previous work from our group (de la Campa, Padilla and de la Cruz, 2017) showed, using a large dataset of neutral and pathogenic variants, that application of this rule has some limitations. However, in this initial analysis the structure of the sequencing experiment was ignored; variants from many different and independent studies were pooled together. In this section, we study the impact of the coincidence rule in the context of a sequencing experiment, using the dataset of 226 panels. To this end, we have applied 15 pathogenicity

predictors present in the dbNSFP database (Liu *et al.*, 2016) (see Materials and Methods) to the variants identified for the 226 individual patients. Then, we explored the coincidence between these predictions, for each variant.

In Figure 5.3a we show the relationship between the number of variants per patient with concordant predictions as a function of the number of predictors employed. We can see that as the latter grows, the number of discrepancies also grows, leading to an increasing number of variants for which *in silico* evidence would be rejected. This result, in accordance with our early work (de la Campa, Padilla and de la Cruz, 2017) constitutes a problem because: (i) some of the predictions are obviously correct, and their use would provide valuable support to the diagnostic process; and (ii) in some cases, loss of a variant also means loss of the carrier gene, which in some cases may correspond to the causal gene.

A potential solution to this problem is the use of metapredictors, which are *in silico* tools combining the results of several pathogenicity predictors (to which we will refer as constituting predictors). Metapredictors, which have flourished in recent years, usually provide an output even when the constituting predictors disagree. Therefore, if properly characterized, metapredictors could complement the coincidence rule in the case where variants have discordant predictions. To explore this issue, we have employed three well-known metapredictors, MetaLR, MetaSVM and REVEL, to annotate the variants in the 226 panels.

In Figure 5.3b we show the gene recovery resulting from the use of these methods, relative to the amount of carrier genes identified with the combination of their constituting predictors. Gene recovery refers to the number of genes that are recovered using metapredictors that were previously lost following the ACMG-AMP guidelines, because the variants

reported in those genes did not match the coincidence rule on the predictions of their constituting predictors. We observe that not all metapredictors are equally powerful: while MetaLR and MetaSVM allow the recovery of between 1 and 10 genes, use of REVEL results in the recovery of between 10 and 30 genes per patients. Thus, while our data indicate that, overall, the use of metapredictors can constitute a good option when the coincidence rule fails, it is clear that the technical details matter and that REVEL is presently the most competitive tool. In summary, we believe that metapredictors can constitute a good option to complement the application of the ACMG-AMP rules for *in silico* evidence, avoiding loss of information.

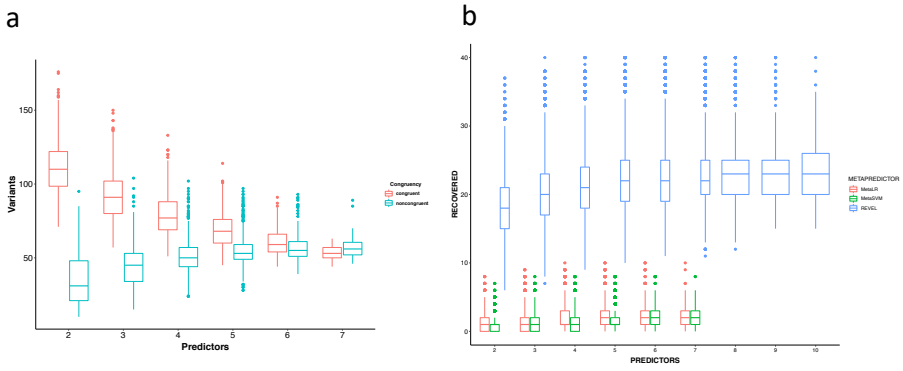


Figure 5.3. Boxplots representing the use of different combinations of in silico pathogenicity predictors. a) *The number of variants per patients that have coincident (blue) and noncoincident (red) predictions. The number of coincident predictions reduces as we combine more predictors, and the number of noncoincident predictions increases.* ***b)*** *The number of genes per patient that we recover when using a metapredictor (MetaLR (red), MetaSVM (green) and REVEL (blue)) instead of using a combination of its constituting predictors.*

5.3.2. Behavior of *in silico* predictors for causal variants

As we have seen before, a total of 24 out 226 patients were diagnosed. For these 24 cases, in 17 of them the causal variant(s) was/were reported to

be missense variant(s). In the following, we focus our analyses on these 17 patients.

For every diagnosed patient between 155 and 175 missense variants were reported (Figure 5.4a); for all of them, we obtained the pathogenicity predictions using the set of 15 tools. Interestingly, we observed (Figure 5.4b) that the 15 causal variants were not always the variants with the maximum number of pathogenicity predictions; some non-causal variants had more.

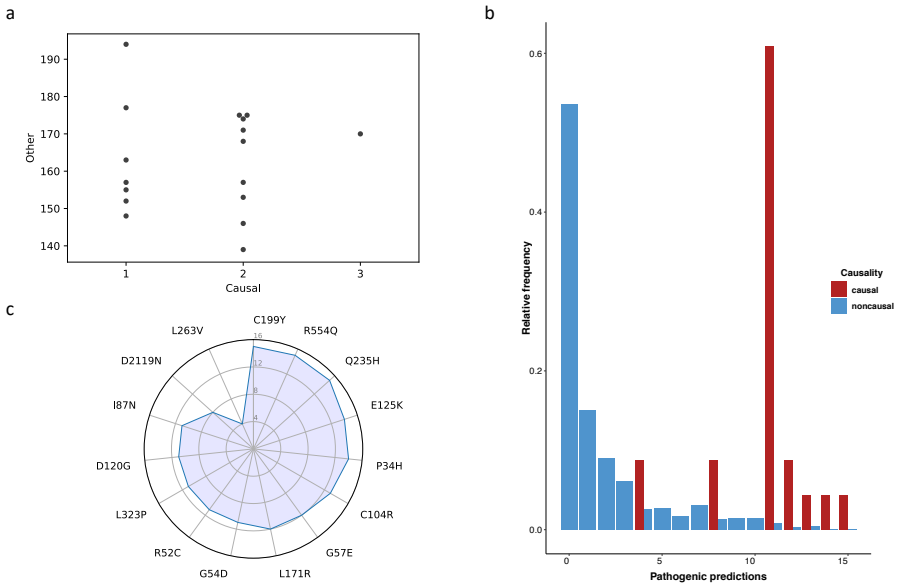


Figure 5.4. Missense variants in diagnosed patients. a) *Distribution of the variants per diagnosed patient.* ***b)*** *Distribution of the variants predicted as pathogenic.* ***c)*** *Radar chart with the number of pathogenic predictions for each causal variant.*

More precisely, we observed that for the causal variants, the number of pathogenic predictions fluctuated between 4 and 15, with only three variants showing a full coincidence between predictors (Figure 5.4c). This means that, when using the predictors in our list, a strict application of the ACMG-AMP guidelines would lead to the loss of most of the variants if different pathogenicity predictors were used to annotate these variants. In the

case of metapredictors, we observed that MetaLR and MetaSVM predicted as pathogenic 9 and 10 of the causal variants, respectively. REVEL correctly predicted as pathogenic the 15 causal variants. Thus, these results reinforce the idea that among the metapredictors REVEL is presently the most competitive tool, based on the data for this panel.

Combining *in silico* evidence and allele frequency

Usually, *in silico* predictions are only part of the evidence utilized to establish a diagnostic. In many cases, when healthcare professionals adhere the ACMG-AMP guidelines, allele frequency is given an important role, when available (Richards *et al.*, 2015). Indeed, allele frequency is utilized as a filtering parameter that can help improve the diagnostic yield, by decreasing the number of candidate variants. When using this parameter, variants with an allele frequency above a certain threshold are discarded; the allele frequencies are usually estimated from databases like the 1000 Genomes Project or GnoMad (Auton *et al.*, 2015; Karczewski *et al.*, 2020). It has to be noted, however, that this filtering step may also have an undesired negative effect, when the decision threshold is set at a too stringent level. For this reason, when adapting ACMG-AMP guidelines to specific diseases, the cutoff level for frequency is one of the adjusted parameters, e.g., see guidelines for the Li-Fraumeni syndrome (*ClinGen TP53 Expert Panel*, 2019), for example.

In this section, we explore this problem, testing how the frequency filtering affects the results of the *in silico* predictors. To this end, for the three selected metapredictors, MetaLR, MetaSVM and REVEL, we display the percentage of causal variants correctly predicted as pathogenic as a function of the threshold applied to the allele frequency. To obtain the different thresholds of the allele frequency, we started with 0.05 and 0.01 values, mentioned in the ACMG-AMP guidelines (Richards *et al.*, 2015), and we

gradually reduced them until we reached the frequency of the prevalence of the disease according to the Orphanet database (Pavan *et al.*, 2017), which was 1-9/100000. In Figure 5.5 we see how increasingly small values of this cutoff result in a drop in the number of causal variants identified. This is a generalized trend for the three predictors, although it becomes more severe when the coverage of the method is already small, as for the non-metapredictors LRT and MutPred (Appendix 2 Figure 9.2.1). Although the reduced sample size of this study does not allow to draw statistically significant conclusions, it nonetheless serves to surface a potential problem in the combination of decision criteria, which could induce the loss of valuable evidence. Interestingly, it also underlines the value of metapredictors methods with large coverage, such as REVEL, an observation coherent with the results of the cost analysis (Chapter 3).

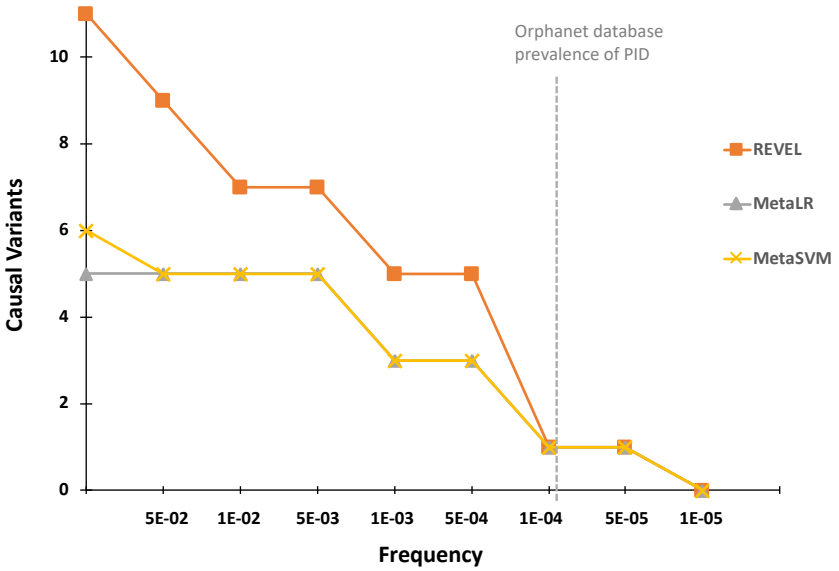


Figure 5.5. Differences in the number of causal variants retrieved when we add the frequency filter. *For three metapredictors, REVEL, MetaLR and MetaSVM, we*

estimated the causal variant retrieval when adding the allele frequency filter. The cutoff values go from 0.05 to the Orphanet database (Pavan *et al.*, 2017) *prevalence.*

5.3.3. Development of a panel-specific *in silico* tool for identifying pathogenic variants

Although the performance of the pathogenicity predictors studied in this chapter is good, it is not enough for their stand-alone usage; they still need improvements. As we have seen in the Introduction, there are different options to improve predictors. One of these is the development of more specific tools (Riera, Padilla and de la Cruz, 2016), adapted to the peculiarities of a given gene or gene family. In this section, we extend this idea to the case of gene panels, presenting the development of a predictor specific for the PID panel. The differential feature of this predictor is that it is trained using only variants for the genes in the panel, retrieved from the UniProt/SwissProt (Bateman *et al.*, 2017) database.

We utilized a Random Forest classifier to build our predictor. In the Table 9.2.1 in Appendix 2 we provide the performance parameters for our tool. We find that its success rate is comparable to, or better than, that of standard pathogenicity predictors (Figure 5.6). It presents the highest MCC (=0.706) and the second highest accuracy (=0.862), just behind MutPred (accuracy=0.892) (Appendix 2 Table 9.2.1, Figure 5.6a). However, the coverage of our RF (100%) is far better than that of MutPred (66.5%) (Figure 5.6b). Regarding sensitivity and specificity, we can see that our RF presents one of the most balanced combination of values, compared to other pathogenicity predictors (Figure 5.6c).

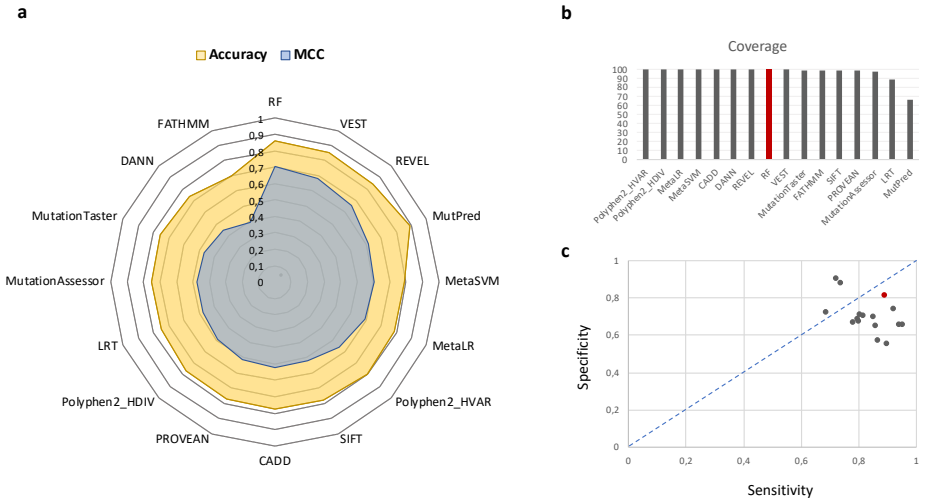


Figure 5.6. Performance estimates of the fifteen pathogenicity predictors and our in-house RF predictor. a) *Radar chart with the accuracy and Matthew's correlation coefficient (MCC) of the 16 methods.* ***b)*** *Bar plot representing the coverage of the 16 methods (RF in red).* ***c)*** *Scatter plot illustrating the positive correlation between sensitivity and specificity for the 16 methods (RF in red).*

Trained in a dataset that did not include any of the causal variants identified in the diagnosed patients, our RF was able to correctly identify 8 of these, out of a total of 15. This result is in the lower range of the results for the majority of predictors, which identified more causal variants (Appendix 2 Figure 9.2.2). These results constitute a proof of concept of the viability of developing specific predictors adapted to designed panels, although in its present form it would only serve to complement present pathogenicity predictors.

5.3.4. Clustering

The results in this chapter show the value of *in silico* predictors for addressing the variant interpretation problem, in the case of the PID panel. The existence of an upper limit in their performance is confirmed by the moderate performance of the previous results showing that even important

changes in the prediction strategy lead to comparable success rates. While improvements in the predictive model may enhance performance, e.g., addition of features based on the protein structure, functional sites, etc., it is also clear that in the case of PID, the lack of a term taking into account background may be a severe limitation.

In this final section, we considered the problem of taking into account the genetic background of the disease, as represented in the panel. We discarded approaches based on the additivity of variants effects (Wells, 1990), of limited reach for the moment. Inspired in the power of recent, non-linear clustering techniques, we decided to use one of them, t-SNE (t-distributed Stochastic Neighbor Embedding) dimensionality reduction algorithm (van der Maaten and Hinton, 2008), using as input features the variability observed for each of the panel genes. Recently employed in the clustering of single-cell populations, t-SNE has shown a good ability to detect hidden relationships in multivariant data. As input to the program we have utilized all the variants identified for an individual to determine his/her state (disease/healthy), thus implementing an approximated version of the idea that the disease is caused by the contribution of the variants in an unknown number of genes. We have utilized healthy individuals from the 1000 Genomes Project (Auton *et al.*, 2015) as a control. We started working in this approach by the end of my Ph.D. work, and the results are preliminar, but given their novelty and potential interest we believe it is valuable to provide a first version of them.

When analyzing the clustering results (Figure 5.7a), we see no differences between the patient and the overall population of healthy individuals. This is particularly the case if we consider only the healthy populations to individuals from the European (EUR) population; these individuals in principle have the same overall ethnic origin as the disease patients. Not surprisingly, the clearest separation happens between the

African (AFR) population and the patients, something to be expected on the basis of independent studies of genetic diversity and disease (Auton *et al.*, 2015; Marín, Aguirre and de la Cruz, 2019). The result was equal performing the analysis using the Principal Components Analysis (PCA) dimensionality reduction algorithm (Hotelling, 1933) (Appendix 2 Figure 9.2.3). In this case, we observed, as mentioned by van der Maaten and Hinton (van der Maaten and Hinton, 2008), an artificial tendency to crowd points together in the center of the map.

The situation changes substantially if, for our analysis, we only take into account the variants predicted as pathogenic by PolyPhen2 and SIFT (directly available data from the 1000 Genomes Project) the previous picture changes substantially. In this case, we observe differences between patient and healthy European populations, indicating that the distribution of missense pathogenic variants in the Primary Immunodeficiency Gene Panel, considered as a whole, differs between healthy individuals and patients (Figure 5.7b). This would be in accordance with the idea of a background contribution to PID, as previously expressed (Gennery, 2016) and opens the way to explore new options to include genetic background in pathogenicity prediction models.

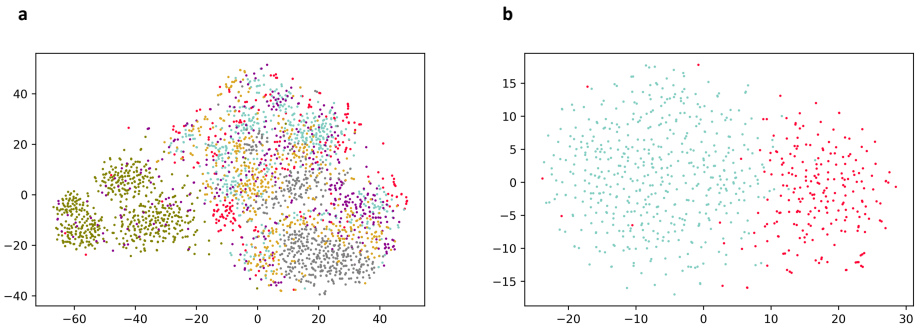


Figure 5.7. tSNE clusterization of the distribution of missense variants in the Primary Immunodeficiency Gene Panel for healthy individuals and patients. a) *The results for the patient population are shown in red, and different colors represent the results for*

each of the five super-populations in the 1000 Genomes Project: African (AFR)(olive-green), Admixed American (AMR)(purple), East Asian (EAS)(grey), European (EUR)(blue) and South Asian (SAS)(golden-yellow) populations. ***b)*** *The results for the comparison of predicted pathogenic variants between European (EUR) (blue) and patients (red) populations.*

5.4. Conclusions

Our results, which are specific for the Primary Immunodeficiency Gene Panel, show that the use of metapredictors can constitute a good option to complement the application of the ACMG-AMP rules for *in silico* evidence, avoiding loss of information. Moreover, among the metapredictors, REVEL has shown to be the most competitive tool, correctly predicting as pathogenic the 15 causal variants. The results obtained from the development of a panel-specific *in silico* tool for identifying pathogenic variants, constitute a proof of concept of the viability of developing specific predictors adapted to designed panels. Finally, although preliminar, the results obtained through the non-linear clustering techniques point in the direction that the genetic background could be modeled using non-linear techniques like t-SNE.

6. General Discussion

The translation of NGS technologies from the research field to the clinical setting (Bertier, Cambon-Thomsen and Joly, 2018) and, specifically, the results obtained in diagnostic yield remain far from expected (Berg, Khoury and Evans, 2011). In this work, we have focused on the "variant interpretation problem" using pathogenicity predictors, a process in which *in silico* tools are used to establish the pathogenic nature of variants explaining the clinical phenotype of their carrier. More precisely, the research developed in this thesis has been devoted to study the components that determine the applicability of *in silico* pathogenicity predictors in the clinical setting. The work describes the efforts we have made in this direction, following three different approaches: (i) developing a cost framework to measure the performance of pathogenicity predictors in the clinical setting (Chapter 3); (ii) studying the role of genetic background as a modulator of the impact of variants (Chapter 4); and (iii) exploring how we can enhance the contribution of *in silico* tools to molecular diagnosis, using the PID gene panel as a model system.

Lessons from estimating the performance of pathogenicity predictors with a cost framework

In the Introduction (Chapter 1) we have described the current performance parameters to measure the success rate (classification/misclassification rates) of pathogenicity predictors, and how they provide an incomplete information when we need to choose these tools for clinical applications. To overcome these limitations, we have developed a cost based framework (Chapter 3) extending the conventional formalism (Adams and Hand, 1999; Drummond and Holte, 2006; Hernández-Orallo, Flach and Ferri, 2012) to adapt it to the used pathogenicity predictors, characterized by a ternary output: either neutral or pathogenic predictions, or no prediction.

Our original results, described in Chapter 3, unveil a different view of pathogenicity predictors in the clinical setting. While using AUC or MCC the slight differences between predictors (Figures 3.9b, 3.10d, 3.11c and 3.12, Chapter 3) make it difficult to decide which one is preferable among the top performers, the use of our model easily identifies an optimal method for the scenario under consideration (Figures 3.9 and 3.10, Chapter 3). Moreover, we have shown that the identity of the optimal method for a specific cost scenario changes depending on ρ, the frequency of pathogenic variants (Figures 3.9c, 3.9d, 3.11a and 3.11b, Chapter 3). These results support the idea that it is important considering both performance and clinical scenario when choosing pathogenicity predictors for clinical applications. In fact, this strategy is implicit in the design of adapted ACMG-AMP guidelines to specific disease (Amendola *et al.*, 2016; Fortuno *et al.*, 2018; *ClinGen TP53 Expert Panel*, 2019), although no quantitative models are used in this case. The relevance of these considerations is underlined by the fact that important factors defining the choice of pathogenicity predictors fluctuate substantially, e.g., the costs of sequencing vary between countries (Schwarze *et al.*, 2018), and also some of the downstream medical decisions and their consequences that depend on budget and drug prices that also change substantially between countries (Barbieri *et al.*, 2005; Smith, Busse and Schreyo, 2008; *Health at a Glance 2019*, 2019; Czech *et al.*, 2020) and within countries (Care, Services and Medicine, 2013).

Lessons from the study of specific systems: the case of FVIII and FIX coagulation factors, and Primary Immunodeficiency (PID) Gene Panel

To date, pathogenicity predictors are based on attributes that only take into account the molecular impact of the variant on the protein sequence the to establish the clinical phenotype (Riera, Lois and de la Cruz, 2014; Niroula and Vihinen, 2016). However, the lack of valuable information on the genetic

background could limit the prediction of a variant effect to those cases where there is a monotonic relationship between the molecular impact and the clinical phenotype. In Chapter 4 of this thesis we explored the extent to which this is the case for CPDs in FVIII and FIX coagulation factors. In the case of CPDs, we know that they tend to have milder molecular effects than non-compensated PDs (Ferrer-Costa, Orozco and de la Cruz, 2007; Barešić *et al.*, 2010) (Figure 4.3), and that compensatory mutations are the principal mechanism suppressing the harmful effects of CPDs (Xu and Zhang, 2014; Jordan *et al.*, 2015). We studied how measures of molecular impact relate to the severity phenotype (Figures 4.3, 4.4). We found that although there is a relation between CPDs and their molecular impact (Figure 4.3), the latter is not strongly related to A and B hemophilia disease severity (Figures 4.4, 9.1.1). This is probably be due to the modulatory effect of genetic background (Pavlova and Oldenburg, 2013), which may be stronger in the case of variants like CPDs. In fact, both computational and experimental studies show that the effect of genetic background plays an important modulatory role in the phenotypic effect of mutations (Breen *et al.*, 2012; Rockah-Shmuel, Tóth-Petróczy and Tawfik, 2015; Vu *et al.*, 2015; Hou *et al.*, 2016; Storz, 2016). Particularly, in the case of hemophilia, genetic background is known to play an important role because variants in hemostasis proteins (other than FVIII and FIX coagulation factors) can modify severity phenotype (Pavlova and Oldenburg, 2013). The study of the compositional properties of genetic background revealed the diversity of the number and kinds of variants in hemostasis proteins (Figures 4.5, 4.6, 4.7), in agreement with the overall trends found in the 1000 Genomes Project (Auton *et al.*, 2015). Moreover, comparable results were observed at the variant level; we found that the same variant may appear with different backgrounds in different individuals (Figure 4.8). Thus, these results confirm that genetic background of FVIII and FIX coagulation factors has the potential to modulate the impact of CPDs. This

could explain clinical observations according to which patients of African origin tend to present more severe versions of the disease (Kruse-Jarres, Barnett and Leissinger, 2008), given their higher proportion of pathogenic variants in hemostasis proteins.

Led by the results obtained in the Chapter 4 of the thesis, we studied the limitations of *in silico* predictors in a Primary Immunodeficiency (PID) Gene Panel (Chapter 5). We described how the use of metapredictors can constitute a good option to complement the application of the ACMG-AMP rules (Richards *et al.*, 2015) for *in silico* evidence, avoiding loss of information (Figure 5.3). The results obtained from the development of the PID panel-specific *in silico* tool for identifying pathogenic variants (Figure 5.6), showed the viability of developing specific predictors adapted to the peculiarities of a given gene family (Riera, Padilla and de la Cruz, 2016). It confirmed, however, the limitations of pathogenicity predictors when relying only on local measures of a variant's impact, ignoring genetic background. In this chapter, we explored, preliminarly, a novel approach to represent the relation between genetic diversity and disease (Auton *et al.*, 2015; Marín, Aguirre and de la Cruz, 2019), based on the use of a recently developed clustering technique (van der Maaten and Hinton, 2008). Although preliminar, the results obtained through the non-linear clustering (Figure 5.7) point in the direction that genetic background not only contributes to PIDs, a previously expressed idea (Gennery, 2016), but also that it can be utilized to discriminate between healthy individuals and disease individuals using a standard, black-box technique.

7. General Conclusions

The main conclusions of the work presented in this thesis, related to the objectives described in the chapter 2, are the following:

1. We have developed a cost framework for assessing and comparing the clinical applicability of pathogenicity predictors, extending the conventional formalism to adapt it to the cases of incomplete coverage.

2. We have unveiled a view of pathogenicity predictors completely different from that obtained when using only standard performance parameters. Our model contributes to take into account factors other than success rate in the election of the best tool for a given application.

3. We have shown that, apart from clinical setting, the nature of the sequenced region as described by ρ, the frequency of pathogenic variants, can play an important role in the choice of an *in silico* tool.

4. We have found that there is a mild relationship between the molecular impact of CPDs in coagulation factors FVIII and FIX and the severity of hemophilias A and B.

5. For FVIII and FIX coagulation factors, we have focused on the contribution of genetic background, showing that it may be relevant when the variants under study have a mild impact on protein function.

6. For the Primary Immunodeficiency (PID) Gene Panel, we have shown that the use of metapredictors can constitute a good option to complement the application of the ACMG-AMP rules for *in silico* evidence. Among these, REVEL has shown to be the most competitive tool.

7. The development of a panel-specific *in silico* tool for identifying pathogenic variants is a proof of concept of the viability of developing specific predictors adapted to designed panels. The performance obtained also confirms the predictive limits of these tools, pointing to the existence of other factors, such as genetic background, that need to be modeled to close the predictive gap.

8. The preliminar results obtained using non-linear clustering techniques indicate that genetic background may contribute to PIDs and show how it can be modeled.

8. Bibliography

Adams, N. M. and Hand, D. J. (1999) 'Comparing classifiers when the misallocation costs are uncertain', *Pattern Recognition*, 32(7), pp. 1139–1147. doi: 10.1067/j.echo.2003.08.013.

Adzhubei, I. A. *et al.* (2010) 'A method and server for predicting damaging missense mutations', *Nature Methods*. Nature Publishing Group, 7(4), pp. 248–249. doi: 10.1038/nmeth0410-248.

Adzhubei, I., Jordan, D. M. and Sunyaev, S. R. (2013) *Predicting functional effect of human missense mutations using PolyPhen-2*, *Current Protocols in Human Genetics*. doi: 10.1002/0471142905.hg0720s76.

Al-Numair, N. S. and Martin, A. C. R. (2013) 'The SAAP pipeline and database: tools to analyze the impact and predict the pathogenicity of mutations.', *BMC genomics*, 14 Suppl 3, p. S4. doi: 10.1186/1471-2164-14-S3-S4.

Altshuler, D. L. *et al.* (2010) 'A map of human genome variation from population-scale sequencing', *Nature*, 467(7319), pp. 1061–1073. doi: 10.1038/nature09534.

Amberger, J. *et al.* (2009) 'McKusick's Online Mendelian Inheritance in Man (OMIM®)', *Nucleic Acids Research*, 37(SUPPL. 1), pp. 793–796. doi: 10.1093/nar/gkn665.

Amendola, L. M. *et al.* (2016) 'Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium', *The American Journal of Human Genetics*, 98(6), pp. 1067–1076. doi: 10.1016/j.ajhg.2016.03.024.

Andrews, T. D., Sjollema, G. and Goodnow, C. C. (2013) 'Understanding the immunological impact of the human mutation explosion', *Trends in Immunology*. Elsevier Ltd, 34(3), pp. 99–106. doi: 10.1016/j.it.2012.12.001.

Angarica, V. E., Orozco, M. and Sancho, J. (2015) 'Exploring the complete mutational space of the LDL receptor LA5 domain using molecular dynamics: Linking snps with disease phenotypes in familial hypercholesterolemia', *Human Molecular Genetics*, 25(6), pp. 1233–1246. doi: 10.1093/hmg/ddw004.

Aspromonte, M. C. *et al.* (2019) *Characterization of intellectual disability and autism comorbidity through gene panel sequencing*, *Human Mutation*. doi: 10.1002/humu.23822.

Auton, A. *et al.* (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), pp. 68–74. doi: 10.1038/nature15393.

Badano, J. L. and Katsanis, N. (2002) 'Beyond mendel: An evolving view of human genetic disease transmission', *Nature Reviews Genetics*, 3, pp. 779–789. doi: 10.1038/nrg910.

De Baets, G. *et al.* (2012) 'SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants.', *Nucleic acids research*, 40(Database issue), pp. D935-9. doi: 10.1093/nar/gkr996.

Baldi, P. *et al.* (2000) 'Assessing the accuracy of prediction algorithms for classification: An overview', *Bioinformatics*, 6(5), pp. 412–424. doi: 10.1093/bioinformatics/16.5.412.

Baldi, P. and Brunak, S. (2001) *Bioinformatics*. 2nd ed. Cambridge, Massachusetts: The MIT Press.

Baldridge, D. *et al.* (2017) 'The Exome Clinic and the role of medical genetics expertise in the interpretation of exome sequencing results', *Genetics in Medicine*. Nature Publishing Group, 19(9), pp. 1040–1048. doi: 10.1038/gim.2016.224.

Barbieri, M. *et al.* (2005) 'Variability of Cost-Effectiveness Estimates for Pharmaceuticals in Western Europe : Lessons for Inferring Generalizability', 8(1), pp. 10–23.

Barešić, A., Hopcroft, Lisa E.M., *et al.* (2010) 'Compensated Pathogenic Deviations: Analysis of Structural Effects', *Journal of Molecular Biology*, 396(1), pp. 19–30. doi: 10.1016/j.jmb.2009.11.002.

Bateman, A. *et al.* (2017) 'UniProt: The universal protein knowledgebase', *Nucleic Acids Research*, 45(Database issue), pp. D158–D169. doi: 10.1093/nar/gkw1099.

Bauce, B. *et al.* (2010) 'Multiple mutations in desmosomal proteins encoding genes in arrhythmogenic right ventricular cardiomyopathy/dysplasia', *Heart Rhythm*, 7, pp. 22–29. doi: 10.1016/j.hrthm.2009.09.070.

de Beer, T. A. P. *et al.* (2013) 'Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset.', *PLoS computational biology*, 9(12), p. e1003382. doi: 10.1371/journal.pcbi.1003382.

Belkadi, A. *et al.* (2015) 'Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants', *Proceedings of the National Academy of Sciences of the United States of America*, 112(17), pp. 5473–5478. doi: 10.1073/pnas.1418631112.

Bell, C. J. *et al.* (2011) 'Carrier testing for severe childhood recessive diseases by next-generation sequencing', *Science Translational Medicine*, 3, p. 65ra4. doi: 10.1126/scitranslmed.3001756.

Berg, J. S., Khoury, M. J. and Evans, J. P. (2011) 'Deploying whole genome sequencing in clinical practice and public health: Meeting the challenge one bin at a time', *Genetics in Medicine*, 13(6), pp. 499–504. doi: 10.1097/GIM.0b013e318220aaba.

de Berg, M. *et al.* (2008) *Computational Geometry: Algorithms and Applications*. 3rd Editio. Edited by Springer. New York.

Bergmann, C. *et al.* (2011) 'Mutations in Multiple PKD Genes May Explain Early and Severe Polycystic Kidney Disease', *Journal of the American Society of Nephrology*, 22, pp. 2047–2056. doi: 10.1681/asn.2010101080.

Bertier, G., Cambon-Thomsen, A. and Joly, Y. (2018) 'Is it research or is it clinical? Revisiting an old frontier through the lens of next-generation sequencing technologies', *European Journal of Medical Genetics*. Elsevier, 61(10), pp. 634–641. doi: 10.1016/j.ejmg.2018.04.009.

Biesecker, L. G. and Green, R. C. (2014) 'Diagnostic clinical genome and exome sequencing', *New England Journal of Medicine*, 370(25), pp. 2418–2425. doi: 10.1056/NEJMra1312543.

Bishop, C. M. (2011) *Pattern Recognition and Machine Learning*. New York: Springer.

Blázquez-Bermejo, C. *et al.* (2019) 'Increased dNTP pools rescue mtDNA depletion in human POLG-deficient fibroblasts', *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 33(6), pp. 7168–7179. doi: 10.1096/fj.201801591R.

BlueShield, R. B. (2020) *Evaluating the Utility of Genetic Panels*.

Bonjoch, L. *et al.* (2019) 'Approaches to functionally validate candidate genetic variants involved in colorectal cancer predisposition', *Molecular Aspects of Medicine*. Elsevier, 69(January), pp. 27–40. doi: 10.1016/j.mam.2019.03.004.

Botstein, D. and Risch, N. (2003) 'Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease.', *Nature genetics*, 33 Suppl(march), pp. 228–37. doi: 10.1038/ng1090.

Boyko, E. J. (1994) 'Ruling out or ruling in disease with the most sensitive or specific diagnostic test: short cut or wrong turn?', *Medical decision making : an international journal of the Society for Medical Decision Making*, 14(2), pp. 175–9. doi: 10.1177/0272989X9401400210.

Breen, M. S. *et al.* (2012) 'Epistasis as the primary factor in molecular evolution', *Nature*, 490(7421), pp. 535–538. doi: 10.1038/nature11510.

Bromberg, Y. and Rost, B. (2007) 'SNAP: Predict effect of non-synonymous polymorphisms on function', *Nucleic Acids Research*, 35(11), pp. 3823–3835. doi: 10.1093/nar/gkm238.

Bromberg, Y., Yachdav, G. and Rost, B. (2008) 'SNAP predicts effect of mutations on protein function.', *Bioinformatics (Oxford, England)*, 24(20), pp. 2397–8. doi: 10.1093/bioinformatics/btn435.

Brunelli, L. *et al.* (2019) 'Targeted gene panel sequencing for the rapid diagnosis of acutely ill infants', *Molecular Genetics and Genomic Medicine*, 7(7), pp. 1–10. doi: 10.1002/mgg3.796.

Camacho, D. M. *et al.* (2018) 'Next-Generation Machine Learning for Biological Networks', *Cell*. Elsevier Inc., 173(7), pp. 1581–1592. doi: 10.1016/j.cell.2018.05.015.

Capriotti, E. and Altman, R. B. (2011) 'Improving the prediction of disease-related variants using protein three-dimensional structure', *BMC Bioinformatics*, 12(S4), p. S3. doi: 10.1186/1471-2105-12-S4-S3.

Care, C. on G. V. in H. C. S. and P. of H.-V., Services, B. on H. C. and Medicine, I. of (2013) *Variation in Health Care Spending*. Edited by J. P. Newhouse et al. Washington (DC): National Academies Press (US).

Care, M. A. *et al.* (2007) 'Deleterious SNP prediction: Be mindful of your training data!', *Bioinformatics*, 23(6), pp. 664–672. doi: 10.1093/bioinformatics/btl649.

Carter, H. *et al.* (2013) 'Identifying Mendelian disease genes with the variant effect scoring tool.', *BMC genomics*, 14 Suppl 3(Suppl 3). doi: 10.1186/1471-2164-14-s3-s3.

Chen, R. *et al.* (2016) 'Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases', *Nature Biotechnology*, 34(5), pp. 531–538. doi: 10.1038/nbt.3514.

Chen, R. and Snyder, M. (2013) 'Promise of personalized omics to precision medicine', *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 5(1), pp. 73–82. doi: 10.1002/wsbm.1198.

Chicco, D. (2017) 'Ten quick tips for machine learning in computational biology', *BioData Mining*. BioData Mining, 10(35), pp. 1–17. doi: 10.1186/s13040-017-0155-3.

Choi, M. *et al.* (2009) 'Genetic diagnosis by whole exome capture and massively parallel DNA sequencing', *Proceedings of the National Academy of Sciences of the United States of America*, 106(45), pp. 19096–19101. doi: 10.1073/pnas.0910672106.

Choi, Y. *et al.* (2012) 'Predicting the functional effect of amino Acid substitutions and indels.', *PloS one*, 7(10), p. e46688. doi: 10.1371/journal.pone.0046688.

Chun, S. and Fay, J. C. (2009) 'Identification of deleterious mutations within three human genomes', *Genome Research*, 19(9), pp. 1553–1561. doi: 10.1101/gr.092619.109.

Cline, M. S. *et al.* (2019) 'Assessment of blind predictions of the clinical significance of BRCA1 and BRCA2 variants', *Human Mutation*, 40, pp. 1546–1556. doi: 10.1002/humu.23861.

ClinGen Hearing Loss Expert Panel Specifications to the ACMG/AMP Variant Interpretation Guidelines Version 1 (2018). Available at: https://www.clinicalgenome.org/affiliation/50007/.

ClinGen TP53 Expert Panel Specifications to the ACMG/AMP Variant Interpretation Guidelines Version 1 (2019). Available at: https://www.clinicalgenome.org/affiliation/50013/.

Colijn, C. *et al.* (2017) 'Toward precision healthcare: Context and mathematical challenges', *Frontiers in Physiology*, 8(MAR), pp. 1–10. doi: 10.3389/fphys.2017.00136.

Collins, F. S. *et al.* (2003) 'A vision for the future of genomics research', *Nature*, 431(April), pp. 835–847.

Cooper, D. N. *et al.* (2013) 'Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease.', *Human genetics*, 132(10), pp. 1077–130. doi: 10.1007/s00439-013-1331-2.

Cover, T. M. and Thomas, J. A. (2006) *Elements of information theory*. Wiley-Interscience.

Craig Venter, J. *et al.* (2001) 'The sequence of the human genome', *Science*, 291(5507), pp. 1304–1351. doi: 10.1126/science.1058040.

Crockett, D. K. *et al.* (2012) 'Consensus : a framework for evaluation of uncertain gene variants in laboratory test reporting Consensus : a framework for evaluation of uncertain gene variants in laboratory test reporting', 48(May).

Cullinane, A. R. *et al.* (2011) 'Homozygosity Mapping and Whole Exome Sequencing to Detect SLC45A2 and G6PC3 Mutations in a Single Patient with Oculocutaneous Albinism and Neutropenia', *Journal of Investigative Dermatology*, 131(10), pp. 139–148. doi: 10.1016/j.physbeh.2017.03.040.

Czech, M. *et al.* (2020) 'A Review of Rare Disease Policies and Orphan Drug Reimbursement Systems in 12 Eurasian Countries', 7(January), p. 416. doi: 10.3389/fpubh.2019.00416.

Dammann, M. and Weber, F. (2012) 'Personalized medicine: Caught between hope, hype and the real world', *Clinics*, 67(SUPPLEMENT), pp. 91–97. doi: 10.6061/clinics/2012(Sup01)16.

David, A. and Sternberg, M. J. E. (2015) 'The Contribution of Missense Mutations in Core and Rim Residues of Protein-Protein Interfaces to Human Disease', *Journal of Molecular Biology*. Academic Press, 427(17), pp. 2886–2898. doi: 10.1016/j.jmb.2015.07.004.

Delaney, S. K. *et al.* (2016) 'Toward clinical genomics in everyday medicine: Perspectives and recommendations', *Expert Review of Molecular Diagnostics*. Taylor & Francis, 16(5), pp. 521–532. doi: 10.1586/14737159.2016.1146593.

DePristo, M. A., Weinreich, D. M. and Hartl, D. L. (2005) 'Missense meanderings in sequence space: a biophysical view of protein evolution.', *Nature reviews. Genetics*, 6(9), pp. 678–687. doi: 10.1038/nrg1672.

Desai, A. N. and Jere, A. (2012) 'Next-generation sequencing: Ready for the clinics?', *Clinical Genetics*, 81(6), pp. 503–510. doi: 10.1111/j.1399-0004.2012.01865.x.

Dong, C. *et al.* (2015) 'Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies', *Human Molecular Genetics*, 24(8), pp. 2125–2137. doi: 10.1093/hmg/ddu733.

Dorfman, R. *et al.* (2010) 'Do common in silico tools predict the clinical consequences of amino-acid substitutions in the CFTR gene?', *Clinical genetics*, 77(5), pp. 464–473. doi: 10.1111/j.1399-0004.2009.01351.x.

Drummond, C. and Holte, R. C. (2000) 'Explicitly representing expected cost: An alternative to ROC representation', *Proceeding of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 198–207.

Drummond, C. and Holte, R. C. (2006) 'Cost curves: An improved method for visualizing classifier performance', *Machine Learning*, 65(1), pp. 95–130. doi: 10.1007/s10994-006-8199-5.

van Durme, J. *et al.* (2011) 'A graphical interface for the FoldX forcefield', *Bioinformatics*, 27(12), pp. 1711–1712. doi: 10.1093/bioinformatics/btr254.

Echave, J. and Wilke, C. O. (2017) 'Biophysical Models of Protein Evolution: Understanding the Patterns of Evolutionary Sequence Divergence', *Annual Review of Biophysics*, 46, pp. 85–103. doi: 10.1146/annurev-biophys-070816-033819.

Edgar, R. C. (2004) 'MUSCLE: Multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Research*, 32(5), pp. 1792–1797. doi: 10.1093/nar/gkh340.

Ernst, C. *et al.* (2018) 'Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics', *BMC Medical Genomics*. BMC Medical Genomics, 11(1), pp. 1–10. doi: 10.1186/s12920-018-0353-y.

Fechter, K. and Porollo, A. (2014) 'MutaCYP: Classification of missense mutations in human cytochromes P450.', *BMC medical genomics*, 7, p. 47. doi: 10.1186/1755-8794-7-47.

Feng, B. J. (2017) 'PERCH: A Unified Framework for Disease Gene Prioritization', *Human Mutation*, 38(3), pp. 243–251. doi: 10.1002/humu.23158.

Fernández-Recio, J. (2011) 'Prediction of protein binding sites and hot spots', *Wiley Interdisciplinary Reviews: Computational Molecular Science*, pp. 680–698. doi: 10.1002/wcms.45.

Ferrer-Costa, C., Orozco, M. and Cruz, X. de la (2007) 'Characterization of Compensated Mutations in Terms of Structural and Physico-Chemical Properties', *Journal of Molecular Biology*, 365(1), pp. 249–256. doi: 10.1016/j.jmb.2006.09.053.

Ferrer-Costa, C., Orozco, M. and de la Cruz, X. (2002) 'Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties.', *Journal of molecular biology*, 315(4), pp. 771–786. doi: 10.1006/jmbi.2001.5255\nS0022283601952556 [pii].

Ferrer-Costa, C., Orozco, M. and de la Cruz, X. (2004) 'Sequence-based prediction of pathological mutations.', *Proteins*, 57(4), pp. 811–9. doi: 10.1002/prot.20252.

Ferrer-Costa, C., Orozco, M. and de la Cruz, X. (2007) 'Characterization of Compensated Mutations in Terms of Structural and Physico-Chemical Properties', *Journal of Molecular Biology*, 365(1), pp. 249–256. doi: 10.1016/j.jmb.2006.09.053.

Ferrer-Costa, C., Orozco, M. and De La Cruz, X. (2002) 'Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties', *Journal of Molecular Biology*. Academic Press, 315(4), pp. 771–786. doi: 10.1006/jmbi.2001.5255.

Fersht, A. (1998) *Structure and mechanism in protein science : a guide to enzyme catalysis and protein folding*. W.H. Freeman.

Flach, P. and Matsubara, E. T. (2008) 'On classification , ranking , and probability estimation', *Probabilistic, Logical and Relational Learning - A Further Synthesis*. Available at: http://drops.dagstuhl.de/opus/volltexte/2008/1382.

Fortuno, C. *et al.* (2018) 'Improved, ACMG-compliant, in silico prediction of pathogenicity for missense substitutions encoded by TP53 variants', *Human Mutation*, 39(8), pp. 1061–1069. doi: 10.1002/humu.23553.

Friedman, J. M. *et al.* (2017) 'Genomic newborn screening: public health policy considerations and recommendations', *BMC Medical Genomics*. BMC Medical Genomics, 10, p. 9. doi: 10.1186/s12920-017-0247-4.

Galano-Frutos, J. J., García-Cebollada, H. and Sancho, J. (2019) 'Molecular dynamics simulations for genetic interpretation in protein coding regions: where we are, where to go and when', *Briefings in Bioinformatics*, 00(July 2019), pp. 1–17. doi: 10.1093/bib/bbz146.

Gennery, A. R. (2016) 'The Evolving Landscape of Primary Immunodeficiencies', *Journal of Clinical Immunology*, 36(4), pp. 339–340. doi: 10.1007/s10875-016-0273-6.

Gerlai, R. (2006) 'Genetic Background', in *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 660–663. doi: 10.1007/3-540-29623-9_2370.

Ghorpade, S. and Limaye, B. V. (2009) *A Course in Multivariate Calculus and Analysis*. New York: Springer.

Ghosh, R., Oak, N. and Plon, S. E. (2017) 'Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines', *Genome Biology*. Genome Biology, 18(1), pp. 1–12. doi: 10.1186/s13059-017-1353-5.

Gilissen, C. *et al.* (2011) 'Unlocking Mendelian disease using exome sequencing', *Genome Biology*, 12(9). doi: 10.1186/gb-2011-12-9-228.

Gilissen, C. *et al.* (2012) 'Disease gene identification strategies for exome sequencing', *European Journal of Human Genetics*. Nature Publishing Group, 20(5), pp. 490–497. doi: 10.1038/ejhg.2011.258.

De Goede, C. *et al.* (2016) 'Role of reverse phenotyping in interpretation of next generation sequencing data and a review of INPP5E related disorders', *European Journal of Paediatric Neurology*. Elsevier Ltd, 20(2), pp. 286–295. doi: 10.1016/j.ejpn.2015.11.012.

Goldfeder, R. L. *et al.* (2016) 'Medical implications of technical accuracy in genome sequencing', *Genome Medicine*. Genome Medicine, 8(1), pp. 1–12. doi: 10.1186/s13073-016-0269-0.

Green, E. D. and Guyer, M. S. (2011) 'Charting a course for genomic medicine from base pairs to bedside', *Nature*, 470(7333), pp. 204–213. doi: 10.1038/nature09764.

Green, R. C. *et al.* (2013) 'ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing', *Genetics in Medicine*, 15(7), pp. 565–574. doi: 10.1038/gim.2013.73.

Green, R. C. *et al.* (2017) 'Corrigendum: ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing (Genetics in Medicine (2013) 15 (565-574) DOI: 10.1038/gim.2013.73)', *Genetics in Medicine*. Nature Publishing Group, 19(5), p. 606. doi: 10.1038/gim.2017.18.

Grimm, D. G. *et al.* (2015) 'The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity', *Human Mutation*, 36(5), pp. 513–523. doi: 10.1002/humu.22768.

Group, S. M. (2015) 'Comprehensive gene panels provide advantages over clinical exome sequencing for Mendelian diseases', *Genome Biology*, 16(1), pp. 134–148. doi: 10.1186/s13059-015-0693-2.

Guerois, R., Nielsen, J. E. and Serrano, L. (2002) 'Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations', *Journal of Molecular Biology*, 320(02), pp. 369–387. doi: 10.1016/S0022-2836(02)00442-4.

Hall, M. A., Moore, J. H. and Ritchie, M. D. (2016) 'Embracing Complex Associations in Common Traits: Critical Considerations for Precision Medicine', *Trends in Genetics*. Elsevier Ltd, 32(8), pp. 470–484. doi: 10.1016/j.tig.2016.06.001.

Han, Y. and He, X. (2016) 'Integrating epigenomics into the understanding of biomedical insight', *Bioinformatics and Biology Insights*, 10, pp. 267–289. doi: 10.4137/BBI.S38427.

Hand, D. J. (2001) 'Measuring diagnostic accuracy of statistical prediction rules', *Statistica Neerlandica*, 55(1), pp. 3–16. doi: 10.1111/1467-9574.00153.

Hand, D. J. (2009) 'Measuring classifier performance: A coherent alternative to the area under the ROC curve', *Machine Learning*, 77(1), pp. 103–123. doi: 10.1007/s10994-009-5119-5.

Hand, D. J. (2010) 'Evaluating diagnostic tests: the area under the ROC curve and the balance of errors.', *Imperial College, London*, pp. 1–18.

Hand, D. J. (2012) 'Assessing the Performance of Classification Methods', *International Statistical Review*, 80(3), pp. 400–414. doi: 10.1111/j.1751-5823.2012.00183.x.

Hand, D. J. and Anagnostopoulos, C. (2013) 'When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance?', *Pattern Recognition Letters*, 34(5), pp. 492–495. doi: 10.1016/j.patrec.2012.12.004.

Hand, D. J. and Anagnostopoulos, C. (2019) 'A better Beta for the H measure of classification performance', *Pattern Recognition Letters*, 40(1), pp. 41–46. doi: 10.1016/j.patrec.2013.12.011.

Hastie, T., Tibshirani, R. and Friedman, J. (2017) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.

Health at a Glance 2019 (2019). OECD (Health at a Glance). doi: 10.1787/4dd50c09-en.

Hehir-Kwa, J. Y. *et al.* (2015) 'Towards a European consensus for reporting incidental findings during clinical NGS testing', *European Journal of Human Genetics*, 23(12), pp. 1601–1606. doi: 10.1038/ejhg.2015.111.

Henikoff, S. and Henikoff, J. G. (1992) 'Amino acid substitution matrices from protein blocks.', *Proceedings of the National Academy of Sciences of the United States of America*, 89(22), pp. 10915–10919. doi: 10.1073/pnas.89.22.10915.

Henikoff, S. and Henikoff, J. G. (1994) 'Position-based sequence weights.', *Journal of molecular biology*, 243(4), pp. 574–8. doi: 10.1016/0022-2836(94)90032-9.

Hernández-Orallo, J., Flach, P. and Ferri, C. (2012) 'A unified view of performance metrics: Translating threshold choice into expected classification loss', *Journal of Machine Learning Research*, 13, pp. 2813–2869.

Hernández-Orallo, J., Flach, P. and Ferri, C. (2013) 'ROC curves in cost space', *Machine Learning*, 93(1), pp. 71–91. doi: 10.1007/s10994-013-5328-9.

Hotelling, H. (1933) 'Analysis of a complex of statistical variables into principal components.', *Journal of Educational Psychology*. US: Warwick & York, 24(6), pp. 417–441. doi: 10.1037/h0071325.

Hou, J. *et al.* (2016) 'The Hidden Complexity of Mendelian Traits across Natural Yeast Populations', *Cell Reports*, 16(4), pp. 1106–1114. doi: 10.1016/j.celrep.2016.06.048.

Hubbard, S. and Thornton, J. M. (1993) 'NACCESS, Computer Program'. London. Available at: http://wolf.bms.umist.ac.uk/naccess/.

Ingram, G. I. C. (1976) 'The history of haemoglobin', *Journal of Clinical Pathology*, 29, pp. 469–479.

Ioannidis, N. M. *et al.* (2016) 'REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants', *American Journal of Human Genetics*. American Society of Human Genetics, 99(4), pp. 877–885. doi: 10.1016/j.ajhg.2016.08.016.

Jamuar, S. S. *et al.* (2016) 'Incidentalome from Genomic Sequencing: A Barrier to Personalized Medicine?', *EBioMedicine*. The Authors, 5, pp. 211–216. doi: 10.1016/j.ebiom.2016.01.030.

Jamuar, S. S. and Tan, E. C. (2015) 'Clinical application of next-generation sequencing for Mendelian diseases', *Human genomics*. Human Genomics, 9, pp. 10–16. doi: 10.1186/s40246-015-0031-5.

Jones, M. A. *et al.* (2013) 'Molecular diagnostic testing for congenital disorders of glycosylation (CDG): Detection rate for single gene testing and next generation sequencing panel testing', *Molecular Genetics and Metabolism*. Elsevier Inc., 110(1–2), pp. 78–85. doi: 10.1016/j.ymgme.2013.05.012.

Jordan, D. M. *et al.* (2011) 'Development and validation of a computational method for assessment of missense variants in hypertrophic cardiomyopathy.', *American journal of human genetics*, 88(2), pp. 183–92. doi: 10.1016/j.ajhg.2011.01.011.

Jordan, D. M., Frangakis, S. G., Golzio, C., Cassa, C. A., *et al.* (2015) 'Identification of cis-suppression of human disease mutations by comparative genomics', *Nature*, 524(7564), pp. 225–229. doi: 10.1038/nature14497.

Kanyongo, G. Y. *et al.* (2007) 'Reliability and Statistical Power: How Measurement Fallibility Affects Power and Required Sample Sizes for Several Parametric and Nonparametric Statistics', *J. Mod. Appl. Stat. Methods*, 6, pp. 81–90.

Karczewski, K. J. *et al.* (2020) 'The mutational constraint spectrum quantified from variation in 141 , 456 humans', *bioRxiv*.

Katsonis, P. and Lichtarge, O. (2014) 'A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness.', *Genome research*, 24(12), pp. 2050–8. doi: 10.1101/gr.176214.114.

Kelly, M. and Semsarian, C. (2009) 'Multiple Mutations in Genetic Cardiovascular Disease. A marker of Disease Severity?', *Circulation: Cardiovascular Genetics*, 2, pp. 182–190. doi: 10.1161/CIRCGENETICS.108.836478.

Kerr, I. D. *et al.* (2017) 'Assessment of in silico protein sequence analysis in the clinical classification of variants in cancer risk genes', *Journal of Community Genetics*. Journal of Community Genetics, 8(2), pp. 87–95. doi: 10.1007/s12687-016-0289-x.

Kim, B. *et al.* (2019) 'Next-generation sequencing with comprehensive bioinformatics analysis facilitates somatic mosaic APC gene mutation detection in patients with familial adenomatous polyposis', *BMC Medical Genomics*. BMC Medical Genomics, 12(1), pp. 1–7. doi: 10.1186/s12920-019-0553-0.

Kim, D. W. *et al.* (2010) 'Whole human exome capture for high-throughput sequencing.', *Genome / National Research Council Canada = Génome / Conseil national de recherches Canada*, 53(7), pp. 568–574. doi: 10.1139/G10-025.

Kimura, M. (1985) 'The role of compensatory neutral mutations in molecular evolution', *Journal of Genetics*. Springer India, 64(1), pp. 7–19. doi: 10.1007/BF02923549.

Kircher, M. *et al.* (2014) 'A general framework for estimating the relative pathogenicity of human genetic variants', *Nature Genetics*. Nature Publishing Group, 46(3), pp. 310–315. doi: 10.1038/ng.2892.

Kitzman, J. O. *et al.* (2015) 'Massively parallel single-amino-acid mutagenesis', *Nature Methods*, 12(3), pp. 203–206. doi: 10.1038/nmeth.3223.

Kleffmann, J. *et al.* (2012) 'Dosage-sensitive network in polycystic kidney and liver disease: Multiple mutations cause severe hepatic and neurological complications', *Journal of Hepatology*, 57, pp. 467–477. doi: 10.1016/j.jhep.2012.03.001.

Knight, J. C. (2009) *Human Genetic Diversity: Functional Consequences for Health and Disease*. Oxford: Oxford University Press.

Ko, J. M. *et al.* (2018) 'A new integrated newborn screening workflow can provide a shortcut to differential diagnosis and confirmation of inherited metabolic diseases', *Yonsei Medical Journal*, 59(5), pp. 652–661. doi: 10.3349/ymj.2018.59.5.652.

Kohane, I., Masys, D. and Altman, R. (2006) 'The Incidentalome: A Threat to Genomic Medicine', *Journal of the American Medical Association*, 296(2), pp. 212–216.

Kohane, I. S., Hsing, M. and Kong, S. W. (2012) 'Taxonomizing, sizing, and overcoming the incidentalome', *Genetics in Medicine*, 14(4), pp. 399–404. doi: 10.1038/gim.2011.68.

Kondrashov, Alexey S., Sunyaev, S. and Kondrashov, F. A. (2002) 'Dobzhansky-Muller incompatibilities in protein evolution', *Proceedings of the National Academy of Sciences of the United States of America*, 99(23), pp. 14878–14883. doi: 10.1073/pnas.232565499.

König, E., Rainer, J. and Domingues, F. S. (2016) 'Computational assessment of feature combinations for pathogenic variant prediction', *Molecular Genetics and Genomic Medicine*, 4(4), pp. 431–446. doi: 10.1002/mgg3.214.

Kruse-Jarres, R., Barnett, B. and Leissinger, C. (2008) 'Immune tolerance induction for the eradication of inhibitors in patients with hemophilia A', *Expert Opinion on Biological Therapy*, 8(12), pp. 1885–1896. doi: 10.1517/14712590802515537.

Ku, C. S. *et al.* (2013) 'A new paradigm emerges from the study of de novo mutations in the context of neurodevelopmental disease', *Molecular Psychiatry*, 18(2), pp. 141–153. doi: 10.1038/mp.2012.58.

Kucukkal, T. G. *et al.* (2015) 'Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins.', *Current opinion in structural biology*, 32, pp. 18–24. doi: 10.1016/j.sbi.2015.01.003.

Kumar, S. *et al.* (2012) 'Evolutionary diagnosis method for variants in personal exomes Neonatal desensitization does not universally prevent xenograft rejection', *Nature Methods*. Nature Publishing Group, 9(9), pp. 855–856. doi: 10.1038/nmeth.2147.

de la Campa, E. álvarez, Padilla, N. and de la Cruz, X. (2017) 'Development of pathogenicity predictors specific for variants that do not comply with clinical guidelines for the use of computational evidence', *BMC Genomics*, 18(Suppl 5), pp. 1–14. doi: 10.1186/s12864-017-3914-0.

Lander, S. *et al.* (2001) 'Initial sequencing and analysis of the human genome International Human Genome Sequencing Consortium* The Sanger Centre: Beijing Genomics Institute/Human Genome Center', *Nature*, 409(February). Available at: www.nature.com.

Landrum, M. J. *et al.* (2014) 'ClinVar: Public archive of relationships among sequence variation and human phenotype', *Nucleic Acids Research*, 42(D1), pp. 980–985. doi: 10.1093/nar/gkt1113.

Landrum, M. J. *et al.* (2016) 'ClinVar: Public archive of interpretations of clinically relevant variants', *Nucleic Acids Research*, 44(D1), pp. D862–D868. doi: 10.1093/nar/gkv1222.

Lee, J. M. (2012) *Axiomatic geometry*.

Lehner, B. (2011) 'Molecular mechanisms of epistasis within and between genes', *Trends in Genetics*. Elsevier Ltd, 27(8), pp. 323–331. doi: 10.1016/j.tig.2011.05.007.

Lek, M. *et al.* (2016) 'Analysis of protein-coding genetic variation in 60,706 humans', *Nature*. Nature Publishing Group, 536(7616), pp. 285–291. doi: 10.1038/nature19057.

Lelieveld, S. H. *et al.* (2015) 'Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions', *Human Mutation*, 36(8), pp. 815–822. doi: 10.1002/humu.22813.

Lemke, J. R. *et al.* (2012) 'Targeted next generation sequencing as a diagnostic tool in epileptic disorders', *Epilepsia*, 53(8), pp. 1387–1398. doi: 10.1111/j.1528-1167.2012.03516.x.

Leong, I. U. S. *et al.* (2015) 'Assessment of the predictive accuracy of five in silico prediction tools, alone or in combination, and two metaservers to classify long QT syndrome gene mutations', *BMC Medical Genetics*. ???, 16(1), pp. 1–13. doi: 10.1186/s12881-015-0176-z.

Li, B. *et al.* (2014) 'In silico comparative characterization of pharmacogenomic missense variants', *BMC Genomics*, 15(Suppl 4), pp. 1–9. doi: 10.1186/1471-2164-15-S4-S4.

Li, J. *et al.* (2018) 'Performance evaluation of pathogenicity-computation methods for missense variants', *Nucleic Acids Research*, 46, pp. 7793–7804. doi: 10.1093/nar/gky678.

Li, T. *et al.* (2013) 'The CDC Hemophilia B mutation project mutation list: a new online resource', *Molecular Genetics & Genomic Medicine*, 1(4), pp. 238–245. doi: 10.1002/mgg3.30.

Linderman, M. D. *et al.* (2014) 'Analytical validation of whole exome and whole genome sequencing for clinical applications', *BMC Medical Genomics*, 7(1), pp. 1–11. doi: 10.1186/1755-8794-7-20.

Liu, X. *et al.* (2016) 'dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs', *Human Mutation*, 37(3), pp. 235–241. doi: 10.1002/humu.22932.

Lohmann, K. and Klein, C. (2014) 'Next Generation Sequencing and the Future of Genetic Diagnosis', *Neurotherapeutics*, 11(4), pp. 699–707. doi: 10.1007/s13311-014-0288-8.

López-Ferrando, V. *et al.* (2017) 'PMut: A web-based tool for the annotation of pathological variants on proteins, 2017 update', *Nucleic Acids Research*, 45(W1), pp. W222–W228. doi: 10.1093/nar/gkx313.

van der Maaten, L. and Hinton, G. (2008) 'Visualizing Data using t-SNE', *Journal of Machine Learning Research*, 9, pp. 2579–2605. doi: 10.1007/s10479-011-0841-3.

MacArthur, D. G. *et al.* (2014) 'Guidelines for investigating causality of sequence variants in human disease', *Nature*, 508, pp. 469–476. doi: 10.1038/nature13127.

Mahmood, K. *et al.* (2017) 'Variant effect prediction tools assessed using independent, functional assay-based datasets: Implications for discovery and diagnostics', *Human Genomics*. Human Genomics, 11(1), pp. 1–8. doi: 10.1186/s40246-017-0104-8.

Majewski, J. *et al.* (2011) 'A new ocular phenotype associated with an unexpected but known systemic disorder and mutation: Novel use of genomic diagnostics and exome sequencing', *Journal of Medical Genetics*, 48(9), pp. 593–596. doi: 10.1136/jmedgenet-2011-100288.

Mardis, E. R. (2010) 'The $ 1 , 000 genome , the $ 100 , 000 analysis ?', *Genome Medicine*, 2(84), pp. 7–9.

Marín, Ò., Aguirre, J. and de la Cruz, X. (2019) 'Compensated pathogenic variants in coagulation factors VIII and IX present complex mapping between molecular impact and hemophilia severity.', *Scientific reports*, 9(1), p. 9538. doi: 10.1038/s41598-019-45916-3.

Marques Matos, C., Alonso, I. and Leão, M. (2019) 'Diagnostic yield of next-generation sequencing applied to neurological disorders', *Journal of Clinical Neuroscience*, 67(xxxx), pp. 14–18. doi: 10.1016/j.jocn.2019.06.041.

Matthijs, G. *et al.* (2016) 'Guidelines for diagnostic next-generation sequencing', *European Journal of Human Genetics*, 24(1), pp. 2–5. doi: 10.1038/ejhg.2015.226.

Miller, M. P. and Kumar, S. (2001) 'Understanding human disease mutations through the use of interspecific genetic variation', *Human molecular genetics*, 10(21), pp. 2319–2328. doi: 10.1093/hmg/10.21.2319.

Miyata, T., Miyazawa, S. and Yasunaga, T. (1979) 'Two types of amino acid substitutions in protein evolution', *Journal of Molecular Evolution*, 12(3), pp. 219–236. doi: 10.1007/BF01732340.

Moles-Fernández, A. *et al.* (2018) 'Computational Tools for Splicing Defect Prediction in Breast/Ovarian Cancer Genes: How Efficient Are They at Predicting RNA Alterations?', *Frontiers in Genetics*, 9, p. 366.

Mook, O. R. F. *et al.* (2013) 'Targeted sequence capture and GS-FLX Titanium sequencing of 23 hypertrophic and dilated cardiomyopathy genes: Implementation into diagnostics', *Journal of Medical Genetics*, 50(9), pp. 614–626. doi: 10.1136/jmedgenet-2012-101231.

Moret, A. *et al.* (2019) 'Next generation sequencing in bleeding disorders: two novel variants in the F5 gene (Valencia-1 and Valencia-2) associated with mild factor V deficiency', *Journal of Thrombosis and Thrombolysis*. Springer US, 48(4), pp. 674–678. doi: 10.1007/s11239-019-01911-z.

Mueller, S. C. *et al.* (2014) 'Pathogenicity prediction of non-synonymous single nucleotide variants in dilated cardiomyopathy', *Briefings in Bioinformatics*, 16(5), pp. 769–779. doi: 10.1093/bib/bbu054.

Muntoni, F. *et al.* (2006) 'Disease severity in dominant Emery Dreifuss is increased by mutations in both emerin and desmin proteins', *Brain*, 129, pp. 1260–1268. doi: 10.1093/brain/awl062.

Navío, D. *et al.* (2019) 'Structural and computational characterization of disease-related mutations involved in protein-protein interfaces', *International Journal of Molecular Sciences*, 20(7). doi: 10.3390/ijms20071583.

Ng, Pauline C and Henikoff, S. (2003) 'SIFT: predicting amino acid changes that affect protein function', *Nucleic Acids Research*, 31(13), pp. 3812–3814. doi: 10.1093/nar/gkg509.

Ng, S. B. *et al.* (2010) 'Exome sequencing identifies the cause of a mendelian disorder', *Nature Genetics*. Nature Publishing Group, 42(1), pp. 30–35. doi: 10.1038/ng.499.

Nguyen, M. T. and Charlebois, K. (2015) 'The clinical utility of whole-exome sequencing in the context of rare diseases-the changing tides of medical practice', *Clinical Genetics*, 88(4), pp. 313–319. doi: 10.1111/cge.12546.

Nijman, I. J. *et al.* (2014) 'Targeted next-generation sequencing: A novel diagnostic tool for primary immunodeficiencies', *Journal of Allergy and Clinical Immunology*. Elsevier Ltd, 133(2), pp. 529-534.e1. doi: 10.1016/j.jaci.2013.08.032.

Niroula, A., Urolagin, S. and Vihinen, M. (2015) 'PON-P2: Prediction method for fast and reliable identification of harmful variants', *PLoS ONE*, 10(2), pp. 1–17. doi: 10.1371/journal.pone.0117380.

Niroula, A. and Vihinen, M. (2016) 'Variation Interpretation Predictors: Principles, Types, Performance, and Choice', *Human Mutation*, 37(6), pp. 579–597. doi: 10.1002/humu.22987.

Nussinov, R. *et al.* (2019) 'Review: Precision medicine and driver mutations: Computational methods, functional assays and conformational principles for interpreting cancer drivers', *PLoS Computational Biology*. Public Library of Science, 15(3). doi: 10.1371/journal.pcbi.1006658.

O'Roak, B. J. *et al.* (2011) 'Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations', *Nature Genetics*, 43(6), pp. 585–589. doi: 10.1038/ng.835.

Olfson, E. *et al.* (2015) 'Identification of medically actionable secondary findings in the 1000 genomes', *PLoS ONE*, 10(9), pp. 1–18. doi: 10.1371/journal.pone.0135193.

Padilla, N. *et al.* (2019) 'BRCA1- and BRCA2-specific in silico tools for variant interpretation in the CAGI 5 ENIGMA challenge.', *Human mutation*, 40(9), pp. 1593–1611. doi: 10.1002/humu.23802.

Pasche, B. and Absher, D. (2011) 'Whole-Genome Sequencing: A Step Closer to Personalized Medicine', *Journal of the American Medical Association*, 305(15), pp. 1596–1597.

Pavan, S. *et al.* (2017) 'Clinical practice guidelines for rare diseases: The orphanet database', *PLoS ONE*, 12(1), pp. 1–14. doi: 10.1371/journal.pone.0170365.

Pavlova, A. and Oldenburg, J. (2013) 'Defining severity of hemophilia: More than factor levels', *Seminars in Thrombosis and Hemostasis*, 39(7), pp. 702–710. doi: 10.1055/s-0033-1354426.

Payne, A. B. *et al.* (2013) 'The CDC Hemophilia A Mutation Project (CHAMP) Mutation List: A New Online Resource', *Human Mutation*, 34(2), pp. E2382–E2391. doi: 10.1002/humu.22247.

Pearson, W. R. (2013) 'Selecting the right similarity-scoring matrix', *Current Protoc. Bioinformatics*, 43, pp. 3.5.1–3.5.9. doi: 10.1002/0471250953.bi0305s43.

Pedregosa, F. *et al.* (2011) 'Scikit-learn : Machine Learning in Python', *Journal of Machine Learning Research*, 12, pp. 2825–2830. doi: 10.1016/j.molcel.2012.08.019.

Pejaver, V. *et al.* (2017) 'MutPred2: inferring the molecular and phenotypic impact of amino acid variants', pp. 1–28. doi: 10.1101/134981.

Pepe, M. S. (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford: Oxford University Press.

Pertea, M. *et al.* (2018) 'CHESS : a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise', *Genome Biology*. Genome Biology, 19, p. 208.

Petukh, M., Kucukkal, T. G. and Alexov, E. (2015) 'On human disease-causing amino acid variants: statistical study of sequence and structural patterns.', *Human mutation*, 36(5), pp. 524–534. doi: 10.1002/humu.22770.

Picard, C. *et al.* (2018) 'International Union of Immunological Societies: 2017 Primary Immunodeficiency Diseases Committee Report on Inborn Errors of Immunity', *Journal of Clinical Immunology*. Journal of Clinical Immunology, 38(1), pp. 96–128. doi: 10.1007/s10875-017-0464-9.

Ponzoni, L. and Bahar, I. (2018) 'Structural dynamics is a determinant of the functional significance of missense variants', *Proceedings of the National Academy of Sciences of the United States of America*, 115(16), pp. 4164–4169. doi: 10.1073/pnas.1715896115.

Quang, D., Chen, Y. and Xie, X. (2015) 'DANN: A deep learning approach for annotating the pathogenicity of genetic variants', *Bioinformatics*, 31(5), pp. 761–763. doi: 10.1093/bioinformatics/btu703.

Randles, L. G. *et al.* (2006) 'Using model proteins to quantify the effects of pathogenic mutations in Ig-like proteins', *Journal of Biological Chemistry*, 281(34), pp. 24216–24226. doi: 10.1074/jbc.M603593200.

Rentzsch, P. *et al.* (2019) 'CADD: Predicting the deleteriousness of variants throughout the human genome', *Nucleic Acids Research*. Oxford University Press, 47(D1), pp. D886–D894. doi: 10.1093/nar/gky1016.

Di Resta, C. *et al.* (2018) 'Next-generation sequencing approach for the diagnosis of human diseases: Open challenges and new opportunities', *Electronic Journal of the International Federation of Clinical Chemistry and Laboratory Medicine*, 29(1), pp. 4–14.

Reva, B., Antipin, Y. and Sander, C. (2011) 'Predicting the functional impact of protein mutations: Application to cancer genomics', *Nucleic Acids Research*, 39(17), pp. 37–43. doi: 10.1093/nar/gkr407.

Ribeiro, Â. M. *et al.* (2015) 'A refined model of the genomic basis for phenotypic variation in vertebrate hemostasis', *BMC Evolutionary Biology*, 15, p. 124. doi: 10.1186/s12862-015-0409-y.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, Wayne W., *et al.* (2015) 'Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology', *Genetics in Medicine*, 17(5), pp. 405–424. doi: 10.1038/gim.2015.30.

Riera, C. *et al.* (2015) 'Molecular damage in Fabry disease: Characterization and prediction of alpha-galactosidase A pathological mutations', *Proteins: Structure, Function and Bioinformatics*, 83(1), pp. 91–104. doi: 10.1002/prot.24708.

Riera, C., Lois, S. and de la Cruz, X. (2014) 'Prediction of pathological mutations in proteins: the challenge of integrating sequence conservation and structure stability principles', *WIREs Computational Molecular Science*, 4, pp. 249–268.

Riera, C., Padilla, N. and de la Cruz, X. (2016) 'The Complementarity Between Protein-Specific and General Pathogenicity Predictors for Amino Acid Substitutions.', *Human mutation*, 37(10), pp. 1013–24. doi: 10.1002/humu.23048.

Rockah-Shmuel, L., Tóth-Petróczy, Á. and Tawfik, D. S. (2015) 'Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations', *PLoS Computational Biology*, 11(8), p. e1004421. doi: 10.1371/journal.pcbi.1004421.

Roden, D. and Tyndale, R. (2013) 'Genomic Medicine, Precision Medicine, Personalized Medicine: What's in a Name?', *Clinical pharmacology and therapeutics*, 94(2), pp. 169–172. doi: 10.1038/jid.2014.371.

Rodrigues, C. *et al.* (2015) 'Performance of In Silico Tools for the Evaluation of UGT1A1 Missense Variants', *Human Mutation*, 36(12), pp. 1215–1225. doi: 10.1002/humu.22903.

Rost, B. and Bromberg, Y. (2009) 'Correlating protein function and stability through the analysis of single amino acid substitutions.', *Bmc Bioinformatics*, 10, p. S8.

Rost, B., Radivojac, P. and Bromberg, Y. (2016) 'Protein function in precision medicine: deep understanding with machine learning', *FEBS Letters*, 590, pp. 2327–2341. doi: 10.1002/1873-3468.12307.

Rudnicki, W. R., Mroczek, T. and Cudek, P. (2014) 'Amino acid properties conserved in molecular evolution', *PLoS ONE*, 9(6), p. e98983. doi: 10.1371/journal.pone.0098983.

Sahni, N. *et al.* (2015) 'Widespread macromolecular interaction perturbations in human genetic disorders', *Cell*. Cell Press, 161(3), pp. 647–660. doi: 10.1016/j.cell.2015.04.013.

Sánchez, I. E. *et al.* (2006) 'Point Mutations in Protein Globular Domains: Contributions from Function, Stability and Misfolding', *Journal of Molecular Biology*, 363, pp. 422–432. doi: 10.1016/j.jmb.2006.08.020.

Sawyer, S. L. *et al.* (2016) 'Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: Time to address gaps in care', *Clinical Genetics*, 89(3), pp. 275–284. doi: 10.1111/cge.12654.

Sboner, A. *et al.* (2011) 'The real cost of sequencing: higher than you think!', *Genome Biology*, 12(8), pp. 125–135. Available at: http://genomebiology.com/2011/12/8/125.

Schwarz, J. M. *et al.* (2014) 'MutationTaster2: mutation prediction for the deep-sequencing age.', *Nature methods*, 11(4), pp. 361–2. doi: 10.1038/nmeth.2890.

Schwarze, K. *et al.* (2018) 'Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature', *Genetics in Medicine*, 20(10), pp. 1122–1130. doi: 10.1038/gim.2017.247.

Schwede, T. (2013) 'Protein modeling: what happened to the "protein structure gap"?', *Structure (London, England : 1993)*, 21(9), pp. 1531–40. doi: 10.1016/j.str.2013.08.007.

Seifi, M. and Walter, M. A. (2018) 'Accurate prediction of functional, structural, and stability changes in PITX2 mutations using in silico bioinformatics algorithms', *PLoS ONE*, 13(4), pp. 1–23. doi: 10.1371/journal.pone.0195971.

Shendure, J., Findlay, G. M. and Snyder, M. W. (2019) 'Genomic Medicine–Progress, Pitfalls, and Promise', *Cell*, 177, pp. 45–57. doi: 10.1016/j.cell.2019.02.003.

Shihab, H. A. *et al.* (2013) 'Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models.', *Human mutation*, 34(1), pp. 57–65. doi: 10.1002/humu.22225.

Shrestha, S. *et al.* (2018) 'Gene-Specific Variant Classifier (DPYD-Varifier) to Identify Deleterious Alleles of Dihydropyrimidine Dehydrogenase.', *Clinical pharmacology and therapeutics*, 104(4), pp. 709–718. doi: 10.1002/cpt.1020.

Sikosek, T. and Chan, H. S. (2014) 'Biophysics of protein evolution and evolutionary protein biophysics.', *Journal of the Royal Society, Interface / the Royal Society*, 11(100), p. 20140419. doi: 10.1098/rsif.2014.0419.

Sim, N. L. *et al.* (2012) 'SIFT web server: Predicting effects of amino acid substitutions on proteins', *Nucleic Acids Research*, 40(W1), pp. 452–457. doi: 10.1093/nar/gks539.

Smith, P. C., Busse, R. and Schreyo, J. (2008) 'VARIABILITY IN HEALTHCARE TREATMENT COSTS AMONGST NINE EU COUNTRIES – RESULTS FROM', *Health Economics*, 17, pp. S1–S8. doi: 10.1002/hec.1330.

Srivastava, A. *et al.* (2013) 'Guidelines for the management of hemophilia', *Haemophilia*, 19(1), pp. e1–e47. doi: 10.1111/j.1365-2516.2012.02909.x.

Starita, L. M. *et al.* (2017) 'Variant Interpretation: Functional Assays to the Rescue', *American Journal of Human Genetics*. American Society of Human Genetics, 101(3), pp. 315–325. doi: 10.1016/j.ajhg.2017.07.014.

Starr, T. N. and Thornton, J. W. (2016) 'Epistasis in protein evolution', *Protein Science*, 25, pp. 1204–1218. doi: 10.1002/pro.2897.

Stassen, J. M., Arnout, J. and Deckmyn, H. (2004) 'The hemostatic system.', *Current med. chem.*, 11, pp. 2245–2260.

Stenson, Peter D. *et al.* (2012) 'The human gene mutation database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution', *Current Protocols in Bioinformatics*, 39, pp. 1.13.1-1.13.20. doi: 10.1002/0471250953.bi0113s39.

Storz, J. F. (2016) 'Causes of molecular convergence and parallelism in protein evolution', *Nat. Rev. Genet.*, 17(4), pp. 239–250. doi: 10.1038/nrg.2016.11.

Stranneheim, H. and Wedell, A. (2016) 'Exome and genome sequencing: A revolution for the discovery and diagnosis of monogenic disorders', *Journal of Internal Medicine*, 279(1), pp. 3–15. doi: 10.1111/joim.12399.

Sun, Y. *et al.* (2019) 'Increased diagnostic yield by reanalysis of data from a hearing loss gene panel', *BMC Medical Genomics*. BMC Medical Genomics, 12(1), pp. 1–8. doi: 10.1186/s12920-019-0531-6.

Sunyaev, S. (2001) 'Prediction of deleterious human alleles', *Human Molecular Genetics*, 10(6), pp. 591–597. doi: 10.1093/hmg/10.6.591.

Sunyaev, S. R. (2012) 'Inferring causality and functional significance of human coding DNA variants.', *Human molecular genetics*, 21(R1), pp. R10-7. doi: 10.1093/hmg/dds385.

Szymanski, M. R. *et al.* (2015) ' Structural basis for processivity and antiviral drug toxicity in human mitochondrial DNA replicase ', *The EMBO Journal*, 34(14), pp. 1959–1970. doi: 10.15252/embj.201591520.

Tang, H. and Thomas, P. D. (2016) 'PANTHER-PSEP: Predicting disease-causing genetic variants using position-specific evolutionary preservation', *Bioinformatics*, 32(14), pp. 2230–2232. doi: 10.1093/bioinformatics/btw222.

Tavtigian, S. V. *et al.* (2006) 'Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral', *Journal of Medical Genetics*, 43(4), pp. 295–305. doi: 10.1136/jmg.2005.033878.

Thomas, P. D. *et al.* (2003) 'PANTHER: A library of protein families and subfamilies indexed by function', *Genome Research*, 13(9), pp. 2129–2141. doi: 10.1101/gr.772403.

Thusberg, J., Olatubosun, A. and Vihinen, M. (2011) 'Performance of mutation pathogenicity prediction methods on missense variants.', *Human mutation*, 32(4), pp. 358–68. doi: 10.1002/humu.21445.

Tian, J. *et al.* (2010) 'Predicting changes in protein thermostability brought about by single- or multi-site mutations', *BMC Bioinformatics*, 11, p. 370. doi: 10.1186/1471-2105-11-370.

To-Figueras, J. *et al.* (2011) 'ALAS2 acts as a modifier gene in patients with congenital erythropoietic porphyria', *Blood*, 118(6), pp. 1443–1451. doi: 10.1182/blood-2011-03-342873.

Tsoutsman, T., Bagnall, R. D. and Semsarian, C. (2008) 'Impact of multiple gene mutations in determining the severity of cardiomyopathy and heart failure', *Clinical and Experimental Pharmacology and Physiology*, 39, pp. 39–49. doi: 10.1111/j.1440-1681.2008.05037.x.

UniProt-Consortium (2014) 'Activities at the Universal Protein Resource (UniProt).', *Nucleic acids research*, 42(Database issue), pp. D191–D198. doi: 10.1093/nar/gkt1140.

Valdar, W. S. J. (2002) 'Scoring residue conservation', *Proteins: Structure, Function and Genetics*, 48(2), pp. 227–241. doi: 10.1002/prot.10146.

Valencia, C. A. *et al.* (2013) 'Comprehensive Mutation Analysis for Congenital Muscular Dystrophy: A Clinical PCR-Based Enrichment and Next-Generation Sequencing Panel', *PLoS ONE*, 8(1), pp. 1–11. doi: 10.1371/journal.pone.0053083.

Vaser, R. *et al.* (2016) 'SIFT missense predictions for genomes', *Nature Protocols*. Nature Publishing Group, 11(1), pp. 1–9. doi: 10.1038/nprot.2015.123.

Velasco, H. and Ramírez-Montaño, D. (2018) 'Incidentalome in neurogenetics: Pathogenic variant of NSD1 in a patient with spinocerebellar ataxia (SCA)', *Frontiers in Genetics*, 9(MAR), pp. 1–5. doi: 10.3389/fgene.2018.00086.

Versteeg, H. H. *et al.* (2013) 'New Fundamentals in Hemostasis', *Physiological Reviews*, 93, pp. 327–358. doi: 10.1152/physrev.00016.2011.

Vihinen, M. (2012a) 'Guidelines for Reporting and Using Prediction Tools for Genetic Variation Analysis', *Human mutation*, 34, pp. 275–282. doi: 10.1002/humu.22253.

Vihinen, M. (2012b) 'How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis.', *BMC genomics*, 13(Suppl. 4), p. S2. doi: 10.1186/1471-2164-13-S4-S2.

Vihinen, M. (2013) 'Guidelines for reporting and using prediction tools for genetic variation analysis.', *Human mutation*, 34(2), pp. 275–82. doi: 10.1002/humu.22253.

Vihinen, M. (2014) 'Majority vote and other problems when using computational tools', *Human Mutation*, 35, pp. 912–914. doi: 10.1002/humu.22600.

Vihinen, M. (2020) 'Problems in variation interpretation guidelines and in their implementation in computational tools', *Molecular Genetics & Genomic Medicine*, 11, p. e1206.

de Visser, J. A. G. M. and Krug, J. (2014) 'Empirical fitness landscapes and the predictability of evolution', *Nature Reviews Genetics*, 15(7), pp. 480–490. doi: 10.1038/nrg3744.

Vu, V. *et al.* (2015) 'Natural Variation in Gene Expression Modulates the Severity of Mutant Phenotypes', *Cell*, 162(2), pp. 391–402. doi: 10.1016/j.cell.2015.06.037.

Wade, N. (2010) 'A Decade Later, Genetic Map Yields Few New Cures', *The New York Times*, p. A1.

Wang, Z. and Moult, J. (2001) 'SNPs, protein structure, and disease', *Human Mutation*, 17(4), pp. 263–270. doi: 10.1002/humu.22.

Weber, S. *et al.* (2016) 'Dealing with the incidental finding of secondary variants by the example of SRNS patients undergoing targeted next-generation sequencing', *Pediatric Nephrology*, 31(1), pp. 73–81. doi: 10.1007/s00467-015-3167-6.

Wells, J. A. (1990) 'Additivity of Mutational Effects in Proteins', *Biochemistry*, 29, pp. 8509–8517. doi: 10.1021/bi00489a001.

Witten, I. H. (Ian H. ., Frank, E. and Hall, M. A. (Mark A. (2011) *Data mining : practical machine learning tools and techniques*. Morgan Kaufmann.

Xu, J. and Zhang, J. (2014) 'Why human disease-associated residues appear as the wild-type in other species: Genome-scale structural evidence for the compensation hypothesis', *Molecular Biology and Evolution*, 31(7), pp. 1787–1792. doi: 10.1093/molbev/msu130.

Xu, S. *et al.* (2017) 'Targeted/exome sequencing identified mutations in ten Chinese patients diagnosed with Noonan syndrome and related disorders', *BMC Medical Genomics*. BMC Medical Genomics, 10(1), pp. 1–7. doi: 10.1186/s12920-017-0298-6.

Xuan, J. *et al.* (2013) 'Next-generation sequencing in the clinic: Promises and challenges', *Cancer Letters*. Elsevier Ireland Ltd, 340(2), pp. 284–295. doi: 10.1016/j.canlet.2012.11.025.

Xue, F. *et al.* (2010) 'Factor VIII gene mutations profile in 148 Chinese hemophilia A subjects', *European Journal of Haematology*, 85(3), pp. 264–272. doi: 10.1111/j.1600-0609.2010.01481.x.

Xue, L. C. *et al.* (2015) 'Computational prediction of protein interfaces: A review of data driven methods', *FEBS Letters*. Elsevier B.V., pp. 3516–3526. doi: 10.1016/j.febslet.2015.10.003.

Xue, Y. *et al.* (2015) 'Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: Single-gene, gene panel, or exome/genome sequencing', *Genetics in Medicine*, 17(6), pp. 444–451. doi: 10.1038/gim.2014.122.

Yaglom, I. M. and Boltyanskii, V. G. (1961) *Convex Figures*. New York: Holt, Rinehart and Winston.

Yates, A. *et al.* (2016) 'Ensembl 2016', *Nucleic Acids Research*, 44(D1), pp. D710–D716. doi: 10.1093/nar/gkv1157.

Yska, H. A. F. *et al.* (2019) 'Diagnostic Yield of Next Generation Sequencing in Genetically Undiagnosed Patients with Primary Immunodeficiencies: a Systematic Review', *Journal of Clinical Immunology*. Journal of Clinical Immunology, 39(6), pp. 577–591. doi: 10.1007/s10875-019-00656-x.

Yu, B. *et al.* (2018) 'Newborn Screening and Molecular Profile of Congenital Hypothyroidism in a Chinese Population', *Frontiers in Genetics*, 9, p. 509. doi: 10.3389/fgene.2018.00509.

Yue, P., Li, Z. and Moult, J. (2005) 'Loss of protein structure stability as a major causative factor in monogenic disease.', *Journal of molecular biology*, 353(2), pp. 459–73. doi: 10.1016/j.jmb.2005.08.020.

Zaghloul, N. A. and Katsanis, N. (2010) 'Functional modules, mutational load and human genetic disease', *Trends in Genetics*, 26, pp. 168–176. doi: 10.1016/j.tig.2010.01.006.

Zaidi, S. *et al.* (2013) 'De novo mutations in histone-modifying genes in congenital heart disease', *Nature*. Nature Publishing Group, 498(7453), pp. 220–223. doi: 10.1038/nature12141.

Zhang, J. and Yang, J.-R. (2015) 'Determinants of the rate of protein sequence evolution.', *Nature reviews. Genetics*, 16(7), pp. 409–420. doi: 10.1038/nrg3950.

Zhang, L. *et al.* (2019) 'Population genomic screening of all young adults in a health-care system: a cost-effectiveness analysis', *Genetics in Medicine*. Springer US, 21(9), pp. 1958–1968. doi: 10.1038/s41436-019-0457-6.

Zhang, X. (2014) 'Exome sequencing greatly expedites the progressive research of Mendelian diseases', *Frontiers of Medicine in China*, 8(1), pp. 42–57. doi: 10.1007/s11684-014-0303-9.

Zhong, Q. *et al.* (2009) 'Edgetic perturbation models of human inherited disorders', *Molecular Systems Biology*, 5, p. 321. doi: 10.1038/msb.2009.80.

Zuckerkandl, E. and Pauling, L. (1962) 'Molecular disease, evolution and genetic heterogeneity', *Horizons in Biochemistry, Academic Press, New York*, pp. 189–225.

9. APPENDICES

Appendix 1:

Additional File 9.1.1. *The complete list set of mild/severe variants used in this work.*

https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-019-45916-3/MediaObjects/41598_2019_45916_MOESM2_ESM.xlsx

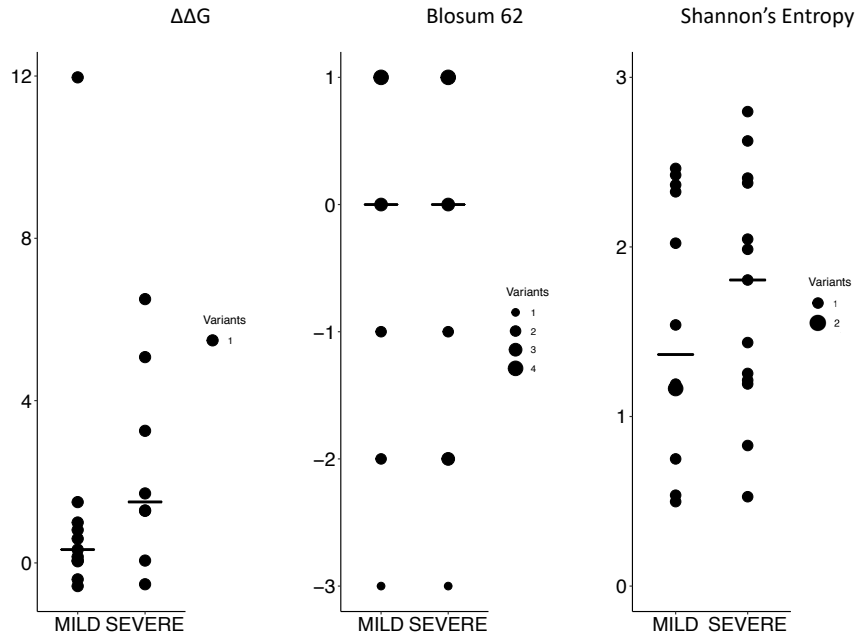


Figure 9.1.1. The molecular impact of CPDs and clinical severity. *For FIX, we plotted the value distribution of three properties (ΔΔG, BLOSUM62 matrix elements and Shannon's entropy) for the severe and mild subsets.*

Table 9.1.1. Causality filtering for the FVIII and FIX CPDs. In the Causality control column: 1 corresponds to variants that passed the filtering, and 0 corresponds to variants that failed the filtering of causality.

Mutation	Uniprot id	Severity	Causality control	Mutation	Uniprot id	Severity	Causality control
A119P	P00451	Mild	0	Q935E	P00451	Severe	1
A1720T	P00451	Mild	1	R1329H	P00451	Mild	1
A1853T	P00451	Severe	1	R1671H	P00451	Severe	1
A194T	P00451	Mild	1	R1740G	P00451	Mild	1
A2066V	P00451	Mild	1	R1768C	P00451	Mild	0
A315V	P00451	Mild	1	R1768H	P00451	Mild	1
A723S	P00451	Mild	1	R2016Q	P00451	Mild	0
D1847G	P00451	Severe	1	R2109C	P00451	Mild	1
D478E	P00451	Mild	1	R2135G	P00451	Mild	1
D963N	P00451	Mild	0	R2169C	P00451	Mild	1
E1057K	P00451	Mild	0	R259G	P00451	Mild	1
E132A	P00451	Mild	1	R259S	P00451	Mild	1
E132D	P00451	Severe	0	R437W	P00451	Mild	0
E1598D	P00451	Severe	0	R458C	P00451	Mild	0
E1863K	P00451	Mild	1	R717Q	P00451	Mild	0
E2006K	P00451	Mild	1	S1530T	P00451	Mild	1
E2200D	P00451	Mild	0	S1807T	P00451	Mild	1
E340K	P00451	Mild	1	S1810P	P00451	Mild	1
E576K	P00451	Mild	0	S1978R	P00451	Severe	0
F1194V	P00451	Mild	1	S202N	P00451	Mild	1
F1794L	P00451	Mild	1	S202R	P00451	Mild	0
F1895S	P00451	Mild	1	T1088I	P00451	Severe	1
F214L	P00451	Mild	1	T137A	P00451	Severe	0
F214V	P00451	Mild	1	T2105I	P00451	Mild	1
F671L	P00451	Mild	0	T2173I	P00451	Mild	0
G164S	P00451	Mild	1	T2173N	P00451	Mild	1
G1729E	P00451	Mild	1	T2272S	P00451	Mild	1
G1769R	P00451	Mild	1	T68A	P00451	Severe	1
G2013R	P00451	Severe	1	T770S	P00451	Mild	1
G2136R	P00451	Severe	1	V115F	P00451	Severe	1
G2344R	P00451	Severe	1	V1981M	P00451	Mild	0
G263D	P00451	Mild	1	V220L	P00451	Mild	1
G434D	P00451	Severe	1	V2251A	P00451	Mild	1
G439C	P00451	Severe	1	V64M	P00451	Mild	1
H1066Y	P00451	Severe	1	V867L	P00451	Severe	1
H113Q	P00451	Severe	1	Y175H	P00451	Mild	0
H113R	P00451	Mild	1	Y450C	P00451	Mild	1
H1234L	P00451	Severe	1	R3H	P00740	Severe	0
H180Y	P00451	Mild	1	I17N	P00740	Severe	0
H1938R	P00451	Mild	1	L20S	P00740	Severe	1
H2174Y	P00451	Mild	1	C28G	P00740	Severe	1
H679Q	P00451	Mild	1	T29I	P00740	Severe	0
I2204T	P00451	Mild	0	N38H	P00740	Severe	1
K146E	P00451	Mild	0	F55I	P00740	Mild	0
K344Q	P00451	Mild	1	G58A	P00740	Severe	1

Table 9.1.1. Continuation.

Mutation	Uniprot id	Severity	Causality control	Mutation	Uniprot id	Severity	Causality control
L107F	P00451	Mild	1	R75Q	P00740	Mild	1
L173R	P00451	Mild	1	P101L	P00740	Mild	0
L1808F	P00451	Mild	0	G105D	P00740	Mild	1
L2249R	P00451	Severe	1	D110G	P00740	Mild	1
L2343P	P00451	Mild	1	C119F	P00740	Severe	1
L296F	P00451	Severe	0	G122R	P00740	Mild	0
L327V	P00451	Mild	1	I136T	P00740	Mild	0
L69V	P00451	Severe	1	E142K	P00740	Severe	0
M166V	P00451	Mild	1	V154A	P00740	Mild	1
M1730V	P00451	Severe	1	S156F	P00740	Severe	1
M1791T	P00451	Severe	0	A164V	P00740	Mild	0
M1842I	P00451	Mild	0	P177S	P00740	Severe	1
M1966V	P00451	Mild	1	F224L	P00740	Severe	0
M1992I	P00451	Mild	1	Q237K	P00740	Severe	1
M1I	P00451	Severe	1	N258K	P00740	Mild	0
M1V	P00451	Severe	1	V277F	P00740	Severe	1
M2218V	P00451	Severe	1	G280R	P00740	Severe	0
M633I	P00451	Mild	1	N283D	P00740	Severe	1
M699T	P00451	Mild	0	Q292K	P00740	Severe	0
M701I	P00451	Mild	0	R294L	P00740	Severe	0
M721V	P00451	Severe	1	E323K	P00740	Mild	0
N1110Y	P00451	Mild	1	L325I	P00740	Mild	0
N1658H	P00451	Mild	1	L325V	P00740	Mild	0
N1913S	P00451	Mild	0	T342A	P00740	Mild	1
N2038S	P00451	Mild	0	I344N	P00740	Mild	0
N2157D	P00451	Mild	1	A366P	P00740	Mild	0
N703D	P00451	Severe	1	R379Q	P00740	Severe	1
P1153A	P00451	Mild	1	A380T	P00740	Mild	1
P1660L	P00451	Severe	1	T381A	P00740	Severe	1
P170S	P00451	Mild	1	K387E	P00740	Mild	1
P1801A	P00451	Mild	1	K387N	P00740	Severe	0
P1844S	P00451	Mild	1	G413E	P00740	Mild	1
P2067L	P00451	Severe	1	H415R	P00740	Severe	1
P2162L	P00451	Severe	1	E420K	P00740	Mild	1
P2311H	P00451	Mild	0	E434G	P00740	Mild	1
Q1705H	P00451	Mild	0	E434K	P00740	Mild	1
Q1764R	P00451	Severe	1	K446N	P00740	Mild	0
Q2208R	P00451	Mild	1	R449Q	P00740	Mild	0
Q2330P	P00451	Severe	1	R449W	P00740	Mild	1
Q324P	P00451	Mild	1				

Table 9.1.2. Set of 19 hemostasis proteins (from Ribeiro et al., BMC Evol Biol, 2015).

Protein Symbol (Ribeiro et al., BMC Evol Biol 2015)	Protein name (Ribeiro et al., BMC Evol Biol, 2015)	UNIPROT Id	Protein Symbol (UniProt)	Number of 1000G males with variants in the gene
vWF	von Willebrand Factor	P04275	vWF	1232
FXII	Coagulation Factor XII	P00748	F12	1225
KLK	Plasma Kallikrein	H0YAC1	KLKB1	1122
ApoH	ApoH, beta2glycoprotein I	P02749	APOH	1060
LRP8	Low density lipoprotein receptor-related protein 8	Q14114	LRP8	914
FII	Coagulation Factor II, Thrombin	P00734	F2	425
PRCP	Prolylcarboxypeptidase or Angiotensinase C	P42785	PRCP	364
GPIbα	Glycoprotein Ib subunit alpha	P07359	GP1BA	343
SERPING1	Serpin G1	E9PGN7	SERPING1	337
FIX	Coagulation Factor IX	P00740	F9	199
GPIX	Glycoprotein IX	P14770	GP9	156
FVIII	Coagulation Factor VIII	P00451	F8	143
GPV	Glycoprotein V	P40197	GP5	117
FX	Coagulation Factor X	P00742	F10	115
FXI	Coagulation Factor XI	P03951	F11	60
CC1Q	Complement Component 1 Binding Protein	Q07021	C1QBP	22
GPIbβ	Glycoprotein Ib subunit beta	P13224	GP1BB	1
HK	Kininogen	P01042	KNG1	5
A2M	Alpha-2-macroglobulin	P01023	A2M	0

Appendix 2:

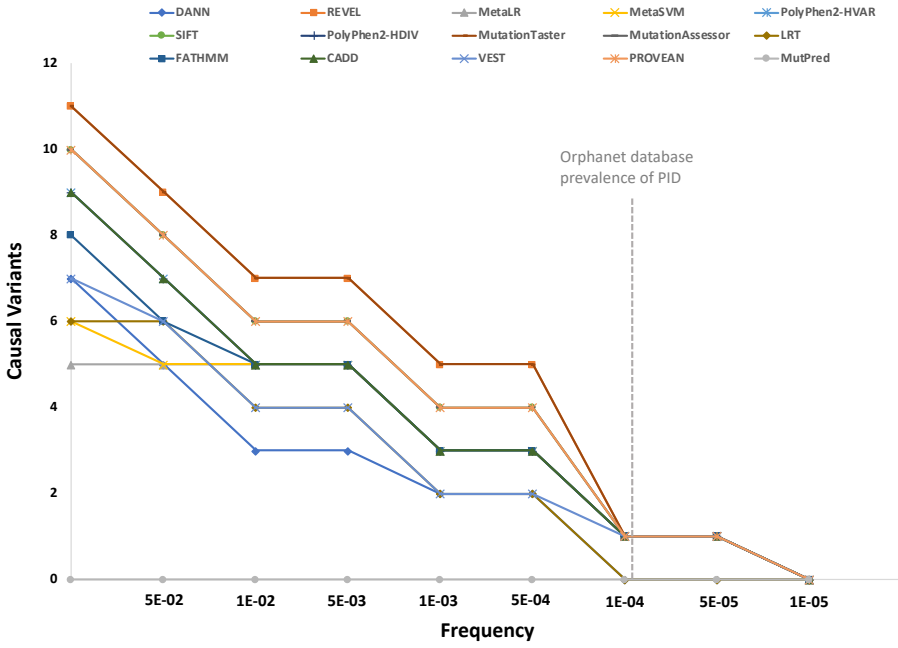


Figure 9.2.1. Differences in the number of causal variants retrieved when we add the frequency filter. *For the 15 pathogenicity predictors in this work, we estimated the causal variant retrieval when adding the allele frequency filter. The cutoff values go from 0.05 to the Orphanet database* (Pavan *et al.*, 2017) *prevalence.*

Table 9.2.1. Performance measures of the fifteen pathogenicity predictors and the Random Forest (RF). *For each predictor, we represent the sensitivity, specificity, PPV, NPV, accuracy, MCC and coverage.*

Predictor	Sensitivity	Specificity	PPV	NPV	Accuracy	MCC	Coverage(%)
Polyphen2-HVAR	0.854	0.696	0.823	0.742	0.795	0.558	100
SIFT	0.861	0.646	0.802	0.736	0.78	0.522	98.485
Polyphen2-HDIV	0.898	0.553	0.769	0.766	0.769	0.492	100
MutationTaster	0.87	0.567	0.773	0.72	0.757	0.464	99.311
MutationAssessor	0.802	0.674	0.8	0.677	0.753	0.476	97.658
LRT	0.797	0.682	0.804	0.673	0.753	0.478	88.843
FATHMM	0.688	0.72	0.805	0.58	0.7	0.396	98.76
MetaLR	0.74	0.875	0.908	0.669	0.791	0.596	100
MetaSVM	0.724	0.901	0.924	0.663	0.791	0.606	100
CADD	0.819	0.703	0.821	0.701	0.775	0.522	100
DANN	0.781	0.667	0.796	0.648	0.738	0.446	100
REVEL	0.956	0.652	0.82	0.899	0.842	0.661	100
VEST	0.923	0.74	0.855	0.852	0.854	0.684	99.862
PROVEAN	0.808	0.705	0.82	0.687	0.769	0.51	98.485
MutPred	0.942	0.655	0.928	0.705	0.892	0.615	66.529
RF	0.894	0.81	0.886	0.822	0.862	0.706	100

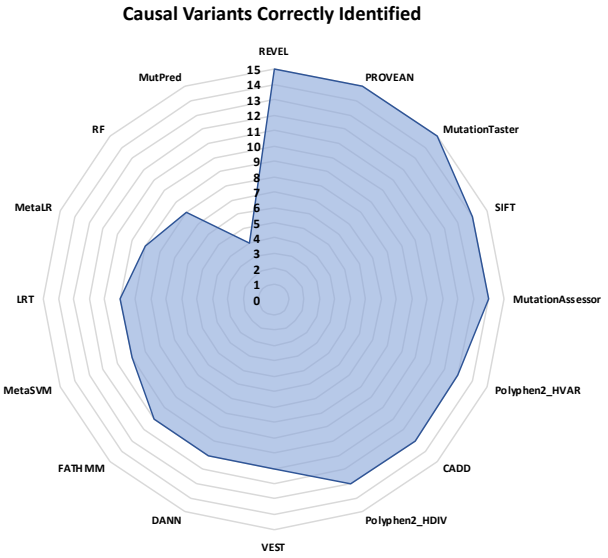


Figure 9.2.2. Radar chart with the causal variants correctly identified by the fifteen pathogenicity predictors and our in-house RF predictor.

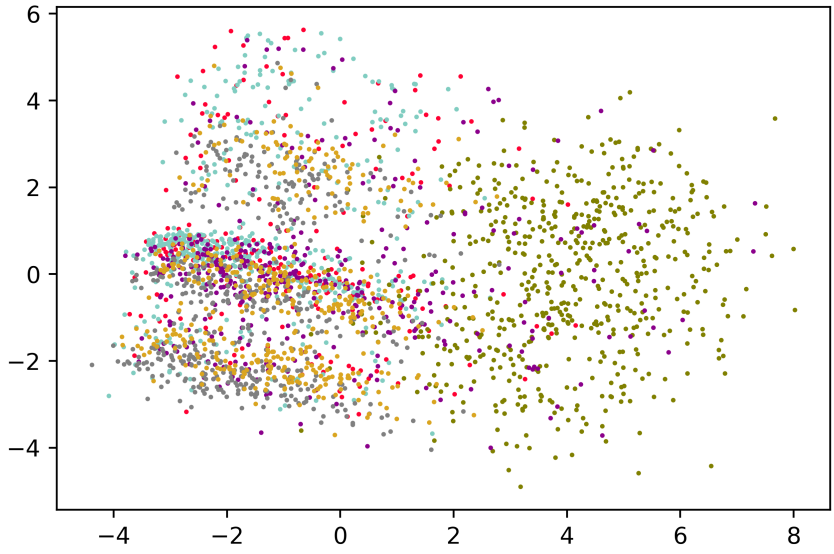


Figure 9.2.3. PCA clusterization of the distribution of missense variants in the Primary Immunodeficiency Gene Panel for healthy individuals and patients. *The results for the patient population are shown in red, and different colors represent the results for each of the five super-populations in the 1000 Genomes Project: African (AFR)(olive-green), Admixed American (AMR)(purple), East Asian (EAS)(grey), European (EUR)(blue) and South Asian (SAS)(golden-yellow) populations.*